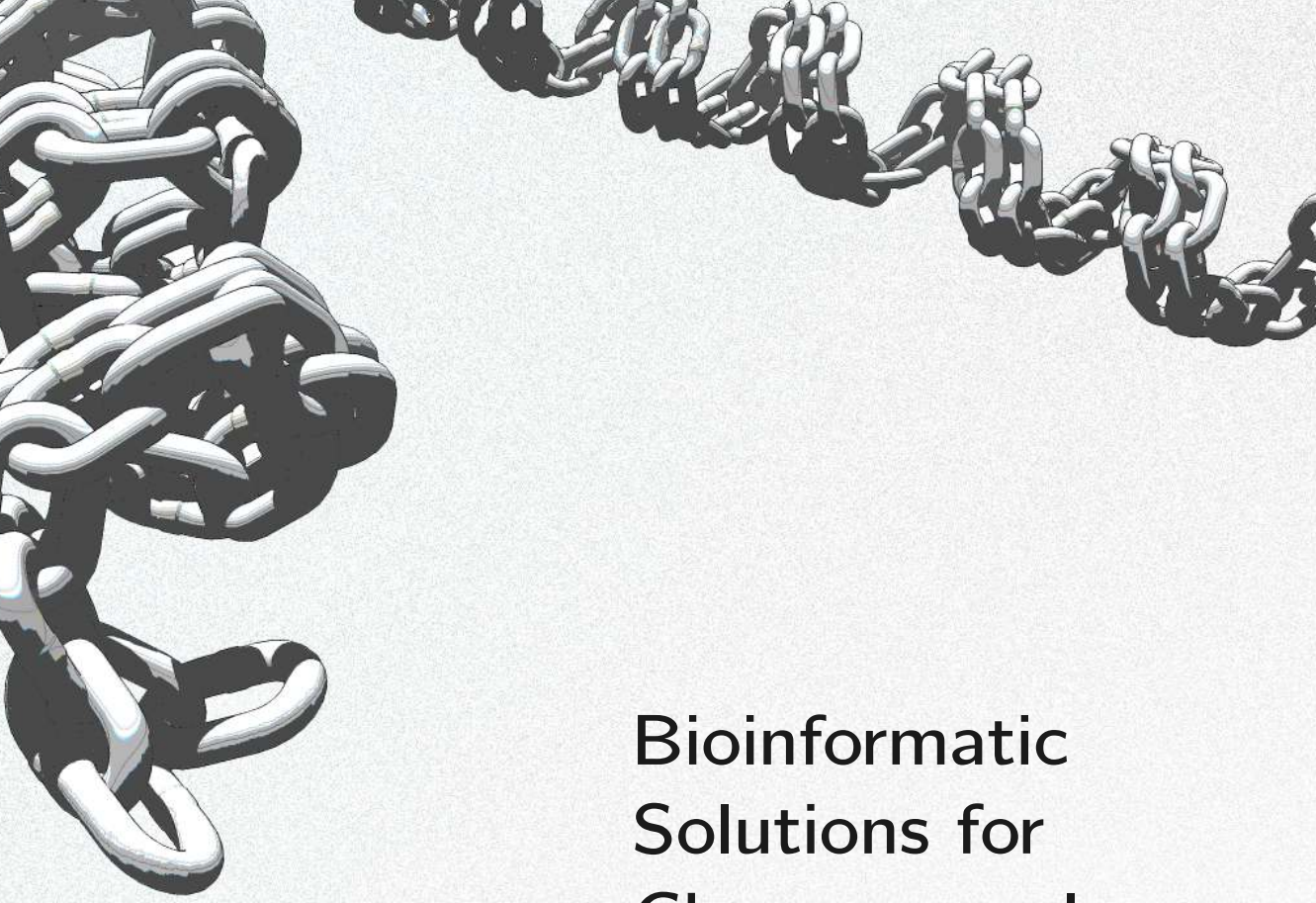

Bioinformatic Solutions for Chromosomal Copy Number Analysis in Cancer

Ilari Scheinin

This page is intentionally left blank.

Bioinformatic solutions for chromosomal copy number analysis in cancer

Jointly-Supervised Academic Dissertation

University of Helsinki, Finland
VU University Amsterdam, The Netherlands

Ilari Scheinin
2017

This is a jointly-supervised (cotutelle) academic dissertation, and the layout is a combination of the conventions of the two universities. The included publications originally published in scientific journals are included as chapters, but in their original layout.

Cover: Henry Scheinin. A chain is only as strong as its weakest link. In translational medicine, this applies to the entire process from the collection and storage of biological specimens, through laboratory and computational techniques, back to the clinic and patient care.

ISBN 978-951-51-3603-9 (paperback)
ISBN 978-951-51-3604-6 (PDF)

Printed by Unigrafia
Helsinki 2017

University of Helsinki

Bioinformatic solutions for chromosomal copy number analysis in cancer

Academic Dissertation

to be presented, with the permission
of the Faculty of Medicine, University of Helsinki,
for public examination in lecture hall 2
of the Haartman Institute, Haartmaninkatu 3,
on Friday October 27, 2017, at 12 noon

Ilari Scheinin

University of Helsinki

Work performed at

Department of Pathology, Medicum
University of Helsinki
Helsinki, Finland

Department of Pathology
VU University Medical Center
Amsterdam, The Netherlands

Supervisors

Professor Sakari Knuutila, PhD
Department of Pathology, Medicum
University of Helsinki
Helsinki, Finland

Professor Bauke Ylstra, PhD
Department of Pathology
VU University Medical Center
Amsterdam, The Netherlands

Reviewers

Professor Matti Nykter, DSc
Faculty of Medicine and Life Sciences
University of Tampere
Tampere, Finland

Research Director, Adjunct Professor Laura Elo, PhD
Turku Centre for Biotechnology
University of Turku
Turku, Finland

Official opponent

Professor Samuel Kaski, PhD
Department of Computer Science
Aalto University
Espoo, Finland

Doctoral Programme in Biomedicine

Vrije Universiteit

Bioinformatic solutions for chromosomal copy number analysis in cancer

Academisch Proefschrift

ter verkrijging van de graad Doctor aan
de Vrije Universiteit Amsterdam,
op gezag van de rector magnificus
prof.dr. V. Subramaniam,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de Faculteit der Geneeskunde
op woensdag 13 december 2017 om 11.45 uur
in de aula van de universiteit,
De Boelelaan 1105

door
Ilari Scheinin
geboren te Helsinki, Finland

promotoren: prof.dr. B. Ylstra

copromotoren: prof.dr. M.A. van de Wiel

Contents

Abbreviations

BAC	bacterial artifical chromosome
BAF	B-allele frequency
CanGEM	Cancer Genome Mine
CBS	circular binary segmentation
CGH	comparative genomic hybridization
CLI	command-line interface
CMG	Laboratory of Cytomolecular Genetics
CNA	copy number aberration
cnLOH	copy-neutral loss of heterozygosity
CNV	copy number variation
DOC	depth of coverage
FISH	fluorescent in-situ hybridization
FFPE	formalin-fixed paraffin-embedded
GEO	Gene Expression Omnibus
GUI	graphical user interface
HMM	hidden Markov model
ICD-10	International Statistical Classification of Diseases and Related Health Problems, 10th Revision
ICD-O-3	International Classification of Disease for Oncology, 3rd Edition
LGG	low-grade glioma
LOH	loss of heterozygosity
MIAME	minimum information about a microarray experiment
miRNA	micro RNA
mRNA	messenger RNA
NGS	next-generation sequencing
PCA	principal component analysis
PCR	polymerase chain reaction
PEM	paired-end mapping
SaaS	software as a service
SNP	single nucleotide polymorphism
SV	structural variant
TNM	TNM classification system for Tumor, lymph Nodes, and Metastasis
WES	whole-exome sequencing
WGS	whole-genome sequencing

Abstract

Chromosomal copy number aberrations are one of the main mechanisms that give rise to the proliferative capabilities of cancer cells. These aberrations can be quantified with technologies that generate measurements genome-wide and with high resolution. Hence, they produce vast amounts of data, which requires tailored bioinformatic solutions for analysis and management. Two such high-resolution and genome-wide technologies are DNA microarrays, which are successively replaced by next-generation sequencing approaches. This dissertation describes three novel bioinformatic solutions for copy number analysis in cancer with these technologies.

CanGEM is a publicly-accessible database solution for storage of raw and processed copy number data from cancer research experiments. The contents of the database can be queried based on clinical and copy number data. Clinical data is collected using appropriate controlled vocabularies. Copy number data is collected as raw microarray data and automated analysis identifies the locations of chromosomal aberrations. In order to allow integration of data measured with different microarray platforms, a copy number status is derived for every known human gene.

CGHpower is a statistical power calculator for copy number experiments that compare two groups. It estimates genome complexity of a cancer type in question from a pilot data set of the sample series, and assesses the number of samples required to satisfy statistical requirements. It can be used either in the planning stages of experiments, including as a justification in grant applications, or to verify whether sufficient samples were included in past experiments. Performance of this bioinformatic solution is evaluated with real and simulated data sets.

QDNAseq is a preprocessing solution to detect copy number aberrations from shallow whole-genome next-generation sequencing data. It corrects the observed sequencing coverage for known systematic biases and allows filtering of spurious regions in the genome. A new list of such problematic regions is derived from public data generated by the 1000 Genomes Project. Performance of the solution is evaluated relative to other similar published solutions and DNA microarrays, and also compared to theoretical statistical expectations.

An application of the QDNAseq method is also presented in a translational research project with the aim to identify copy number aberrations in tumors of patients with low-grade glioma. Aberrations identified by shallow whole-genome next-generation sequencing and QDNAseq are used to evaluate associations with patient survival, and also to assess intratumoral heterogeneity and temporal evolution of these tumors. A loss in chromosome 10q is identified to be associated with poor prognosis, and the finding validated in two independent data sets. From the assessment of intratumoral heterogeneity and temporal tumor evolution, the well-characterized co-deletion of 1p/19q is found to be the only chromosomal aberration that is consistently present or absent across the entire tumor and possible future recurrences. This is compatible with the present view of its role as an early event in the development of these tumors.

The text concludes with a discussion of lessons learned from the development process and application of the three described bioinformatic solutions. Better awareness of and adherence to established best practices from the software development field would have been useful, and together with more careful consideration of implementation decisions could have resulted in software that was more apt for its purpose while also more efficient to develop and maintain. Similar to the presented solutions, much of the development of custom bioinformatics software is performed within academic research groups. Closer attention to the software development process itself could possibly be beneficial for academic software development in general.

Tiivistelmä

Kromosomaaliset kopiolukupoikkeamat ovat eräs tärkeimmistä mekanismeista syövän synnyssä. Yhden äidiltä ja yhden isältä perityn geenikopion sijaan osa perimästä voi olla monistunut useammaksi kopioksi, ja joidenkin osien kohdalla yksi tai molemmat kopiot voivat olla hävinneet. Kopiolukupoikkeamien todentamiseen käytetään genominlaajuisia tekniikoita, joilla on tarkka erotuskyky. Ne tuottavat suuria tietomääriä, joiden analysointi ja käsittely vaativat räätälöityjä bioinformaattisia menetelmiä. Tekniikoihin sisältyvät DNA-mikrolevyt sekä ne käytännössä jo syrjäyttäneet uuden sukupolven sekvensointimenetelmät. Tässä väitöskirjassa kuvataan kolme uutta bioinformaattista ohjelmistoa kopiolukupoikkeamien analysointiin syöpänäytteistä näillä tekniikoilla.

CanGEM on julkinen tietokanta raa'an ja prosessoidun mikrolevyaineiston keräämiseen yksittäisistä syöpätutkimuksista. Tietokannan sisältöön voi tehdä hakuja kliinisten muuttujien tai kopiolukupoikkeamien perusteella. Kliinisten muuttujien tallennukseen käytetään asianmukaisia luokittelujärjestelmiä. Kopiolukuaineisto kerätään raakoina mikrolevymittauksina, joista kopiolukupoikkeamat tunnistetaan algoritmisesti. Jotta eri mikrolevyalustoilla mitatun tiedon yhdistäminen olisi mahdollista, kopioluku määritetään erikseen jokaiselle tunnetulle ihmisen geenille.

CGHpower on menetelmä tilastollisten voima-analyysien tekemiseen kahta ryhmää vertailevista kopiolukututkimuksista. Aineiston kopiolukupoikkeamien monimutkaisuus arvioidaan koe-erästä näytteitä ja määritetään tilastollisten vaatimusten edellyttämä otoskoko. Menetelmää voidaan käyttää joko tutkimusten suunnitteluvaiheessa, mm. rahoitushakemusten tukena, tai arvioimaan onko jo tehdyissä kokeissa käytetty riittävää määrää näytteitä. Suorituskyky mitataan sekä todellisilla että simuloiduilla aineistoilla.

QDNAseq on esikäsittelymenetelmä kopiolukupoikkeamien tunnistamiseen matalalla lukupeitolla ja genominlaajuisesti tuotetusta uuden sukupolven sekvensointiaineistosta. Se korjaa havaittua lukupeittoa tunnettujen vinoumalähteiden osalta ja mahdollistaa kopiolukuanalyyseille ongelmallisten perimän osien suodattamisen jatkokäsittelystä. Näistä ongelmallisista alueista kuvataan uusi luettelo, joka on johdettu 1000 Genomes -projektin julkaisemasta aineistosta. Menetelmän suorituskykyä arvioidaan verrattuna muihin vastaaviin julkaistuihin menetelmiin ja DNA-mikrolevyihin, sekä suhteessa teoreettisiin tilastollisiin odotuksiin.

Itse menetelmän lisäksi kuvataan QDNAseq:n sovellutus translationaaliseen tutkimukseen ja kopiolukupoikkeamien tunnistamiseen alhaisen erilaistumisasteen glioomista. Todetaan kromosomin 10q häviämän yhteys huonoon ennusteeseen ja löydös vahvistetaan kahdessa riippumattomassa aineistossa. Tunnistettuja kopiolukupoikkeamia käytetään myös kasvaimien epäyhtenäisyyden ja ajallisen kehityksen tarkasteluun. Havaitaan kyseiselle syöpätyypille yleisen 1p/19q-häviämän olevan ainoa kopiolukupoikkeama, joka on johdonmukaisesti joko läsnä taikka puuttuu läpi sekä koko alkuperäisen syöpäkasvaimen että mahdollisten uusiutumien. Havainto sopii nykynäkemykseen kyseisen poikkeaman synnystä hyvin varhaisessa vaiheessa kyseisen syöpätyypin kehitystä.

Lopuksi tarkastellaan kuvattujen bioinformaattisten ohjelmistojen kehitys- ja sovellutusprosesseista opittuja asioita. Ohjelmistokehitysalan vakiintuneiden käytänteiden parempi tuntemus olisi ollut hyödyllistä, ja yhdessä toteutusyksityiskohtien tarkemman harkinnan kanssa voinut auttaa tuottamaan tarkoituksensa paremmin täyttäviä sekä helpommin kehitettäviä ja ylläpidettäviä ohjelmistoja. Kuten tässä kuvatut, suuri osa räätälöidyistä bioinformaattisista ohjelmistoista kehitetään akateemisissa tutkimusryhmissä. Suurempi panostus itse ohjelmistokehitysprosessiin voisi yleisesti ottaen hyödyttää akateemista ohjelmistokehitystä.

Samenvatting

Afwijkingen in het aantal chromosomen, of delen van chromosomen, zijn een van de mechanismen die aanleiding geven tot het proliferatieve gedrag van kankercellen. Deze chromosomale afwijkingen kunnen worden gemeten met genomische technieken met een hoge resolutie. Deze technieken genereren zeer grote hoeveelheden data, die op maat gemaakte bioinformatische oplossingen vereisen voor analyse en databeheer. De twee meest relevante genomische technieken met hoge resolutie zijn microarrays en 'next generation sequencing'. Hoofdstuk 1 van dit proefschrift behandelt de literatuur van de data-analyse voor chromosomale afwijkingen gemeten met microarrays of 'next generation sequencing'. Het introduceert relevante bioinformatische concepten, beschrijft het analytische proces van ruwe data tot identificatie van numerieke chromosoomafwijkingen in individuele tumoren en het bioinformatisch onderzoek gericht op de betekenis van die afwijkingen in grote series tumoren.

Hoofdstuk 2 tot en met 4 beschrijven drie nieuwe bioinformatische implementaties ontwikkeld voor de analyse van deze chromosomale afwijkingen in kanker. CanGEM (Hoofdstuk 2) is een publiek toegankelijke database voor het opslaan van ruwe en verwerkte chromosoomaantallen het kankeronderzoek. De inhoud van de database kan worden doorzocht op basis van zowel klinische als experimentele gegevens met betrekking tot chromosoomaantallen. Klinische gegevens worden verzameld met behulp van gecontroleerde woordenlijsten. Chromosoomaantallen worden verzameld als ruwe microarray data en begin- en eindpositie van de afwijkingen worden steeds opnieuw automatisch bepaald. Om de integratie van de data, die gemeten worden met microarrays van verschillende makelij, verder te faciliteren, wordt het aantal chromosomen per gen afgeleid voor ieder van de ca. 19.000 tot 20.000 menselijke genen.

CGHpower (Hoofdstuk 3) is een methode om te berekenen hoeveel tumormonsters statistisch nodig zijn om verschillen en overeenkomsten in chromosomale afwijkingen tussen twee groepen tumoren te kunnen vergelijken. Er wordt een schatting gemaakt van de complexiteit van de afwijkingen in een bepaald type kanker met behulp van een beperkt aantal monsters. Vervolgens wordt geschat hoeveel tumoren nodig zijn om aan de statistische eisen te voldoen. CGHpower kan in de planningsfase van een subsidieaanvraag worden gebruikt als rechtvaardiging van de voorgestelde aantallen naar een subsidiegever, of kan gebruikt worden om te controleren of er voldoende aantallen tumoren in een experiment werden opgenomen. CGHpower wordt geëvalueerd met behulp van experimentele en gesimuleerde datasets.

QDNAseq (Hoofdstuk 4) is een methode die een voorbewerkingstap maakt van 'next generation sequencing' data naar chromosoomaantallen in het genoom van een tumor, waarbij wordt uitgegaan van sequencing met een diepte van slechts 10% van het gehele genoom. QDNAseq corrigeert de waargenomen genoomwijde dekking voor systematische fouten en faciliteert de mogelijkheid om onregelmatige gebieden in het genoom te verwijderen. Een lijst van dergelijke systematische fouten en onregelmatige gebieden is afgeleid van publieke data die openbaar werd gemaakt door het "1000 Genomes Project". QDNAseq wordt geëvalueerd ten opzichte van de microarraytechniek en andere gepubliceerde software voor de analyse van numerieke chromosoomafwijkingen met behulp van 'next generation sequencing'. Tenslotte worden de uitkomsten van QDNAseq op 'next generation sequencing' data vergeleken met theoretische statistisch verwachte resultaten.

In het voorlaatste hoofdstuk (Chapter 5) wordt QDNAseq toegepast op translationeel onderzoek dat tot doel heeft afwijkingen in het aantal chromosomen of delen daarvan te

identificeren bij tumoren van patiënten met laag-gradige gliomen. Chromosomale afwijkingen geïdentificeerd middels 'next generation sequencing' en QDNAseq worden gebruikt om associaties te bepalen met de overleving van de patiënt, de intratumorale heterogeniteit van de tumoren en de evolutie over tijd van deze tumoren. Een verlies van het distale deel van chromosoom 10q wordt in dit onderzoek geassocieerd met een slechte prognose. Deze bevinding kon worden gevalideerd in twee onafhankelijke patiëntenseries. Uit de beoordeling van intratumorale heterogeniteit en tumorevolutie blijkt tenslotte dat verlies van chromosoom 1p samen met 19q de enige afwijking is die consistent aan- of afwezig is in de tumoren.

Net als bij de drie beschreven implementaties voor de analyse van chromosomale afwijkingen in kanker, wordt veel bioinformatisch onderzoek uitgevoerd in academische groepen. De discussie (Hoofdstuk 6) behandelt de opgedane ervaringen met betrekking tot het ontwikkelingsproces en de toepassing van bioinformatische oplossingen.

Chapter 1

Introduction

Chromosomal aberrations in cancer

Cancer is a disease in which control of the cell cycle fails, leading to cell proliferation. This is caused by aberrant signal processes within and between cells. Perturbations in signaling networks allow cancer cells to obtain their characteristic traits, which include sustained proliferative signaling, evasion of growth suppressors and cell death, replicative immortality, induction of angiogenesis, and activation of invasion and metastasis (Hanahan and Weinberg, 2011).

One of the mechanisms through which cells acquire the necessary alterations is chromosomal instability, which includes alteration of chromosomal copy numbers. Compared to the normal diploid copy number of two, the presence of additional copies, or loss of one or both copies, can result in tumorigenic alterations in cell homeostasis (Albertson et al., 2003; Beroukhim et al., 2010). Alternate chromosomal copy numbers can be acquired (somatically), or be present in the germ-line (hereditary). Throughout this text, germ-line alterations are referred to as copy number variations (CNVs) (Church et al., 2010; 1000 Genomes Project Consortium et al., 2015). Some CNVs have been shown to affect cancer risk (Shlien and Malkin, 2010). The focus of this dissertation, however, is on acquired (or somatic) events, which are referred to as chromosomal copy number aberrations (CNAs). To discuss both CNAs and CNVs at the same time, the term copy number alterations is used.

Recurrent CNAs in cancer tissues have a high probability to contain genes whose function is crucial to normal cell homeostasis and perturbations therefore tumorigenic (Beroukhim et al., 2007). Many such genes are affected, directly or indirectly, by CNAs (Futreal et al., 2004; Cox et al., 2005; Beroukhim et al., 2010; Santarius et al., 2010; Kim et al., 2013b). Within tumors localized in the same organ, CNA patterns have been shown to define subtypes (Fridlyand et al., 2006; Chin et al., 2007a; Jong et al., 2007; Habermann et al., 2009; Russnes et al., 2010). Within these subtypes, prognostics and treatment responses can be very different (Curtis et al., 2012; Dancey et al., 2012), and accurate identification of subtypes can thus be of great clinical value (Macintyre et al., 2016). Specific CNAs have also been linked to different environmental risk factors, such as asbestos-exposure (Nymark et al., 2006) or smoking (Dumanski et al., 2014). Meta-analyses of various cancer types have shown that the observed CNA patterns differ between cancers whose precursor cells originate from different germ layer lineages in the embryo (Myllykangas et al., 2006; Jong et al., 2007; Hoadley et al., 2014).

In order for CNAs to be tumorigenic, they would need to affect cell signaling pathways. These pathways are guided by proteins. Proteins are translated from messenger-RNAs (mRNAs), which in turn are transcribed from DNA. The cancerous effect of a CNA would therefore need to be reflected first on the RNA level, and then on a protein level, resulting either in changes in concentration, or molecular modifications such as different configurations, phosphorylations, and glycosylations. Such effects from CNAs can be direct (higher level of transcription and translation for a gene located within an amplification), or indirect, such as CNAs affecting promoter regions, micro-RNAs, or transcription factors or other regulatory proteins.

Both protein (Zhang et al., 2016) and RNA expression (Curtis et al., 2012) levels show correlation with DNA copy number. In fact, a proportion of differential RNA expression in colorectal cancer can be explained in terms of underlying CNAs (Tsafrir et al., 2006). CNAs that encompass oncogenes or tumor-suppressor genes can therefore have a direct effect on cell homeostasis.

Some RNAs, such as micro-RNAs (miRNAs), function as regulators of translation or

degradation of mRNAs. A single miRNA can control hundreds of protein-coding genes, and CNAs that contain miRNAs therefore have tumorigenic effects (Varambally et al., 2008; Croce, 2009). Other possible mechanisms of action include CNAs encompassing transcription factor binding sites, or regions responsible for maintaining the three-dimensional structure of chromatin (Gilbert and Allan, 2014).

As most functional consequences of CNAs are probably carried out via RNA and proteins, one may wonder, why not study those directly instead of DNA? There are both experimental and biological advantages and disadvantages to any class of biomolecules (Smeets et al., 2011). Compared to proteins, the ability of nucleotide sequences to form double helices is a big advantage from the experimental point of view. It allows one to use single-stranded DNA molecules and measure events involving the specific complementary nucleotide sequence.

This applies similarly to both DNA and RNA, and both have been used extensively as targets for diagnostic and experimental techniques in cancer research. Gene expression profiles have been shown to have a strong prognostic value for a wide range of malignancies including leukemias as well as bladder, breast, esophageal, non-small cell lung, and head and neck cancers (Bell, 2010).

Compared to RNA, one of the main advantages of DNA is its stability (Smeets et al., 2011). For more than a century, pathology institutes around the world have been routinely collecting specimens in their archives, usually as formalin-fixed paraffin-embedded (FFPE) blocks (Blow, 2007; Casparie et al., 2007). Together with the accompanying clinical data, these archives present a huge resource for cancer research and biomarker discovery. For studies that require long clinical followup, they may be the only available source of material.

However, while FFPE-fixation works well for histological purposes, not all biomolecules preserve well. This is especially true for RNA, which is often differentially degraded in archival samples, obscuring expression analytical procedures (Scicchitano et al., 2006; Frank et al., 2007; Fedorowicz et al., 2009). DNA is more stable, and therefore a more robust target for diagnostics and studies on such material. Also within DNA techniques, some are more sensitive, while others are more tolerant to degradation (Krijgsman et al., 2012). Since the ability to use archival material is an important aspect in cancer research, compatibility of techniques with DNA isolated from FFPE is a valuable criterion when choosing which techniques to use.

Challenges for data analysis of CNAs

Prior to reviewing analytical techniques for CNAs in the next section, some aspects of their biology are discussed here. The focus is on those features that have a direct impact on interpretation of data from high-throughput analytical methods. Aspects that do not, such as the mechanisms of their genesis, are omitted.

Aberration length and magnitude

The size of CNAs varies from entire chromosomes and chromosome arms to focal aberrations (Krijgsman et al., 2014). As technologies have improved, the spatial resolution to detect CNAs has also increased, allowing ever smaller events to be detected.

CNAs manifest as a range of chromosomal copy numbers that deviate from the diploid copy number in healthy somatic cells. Gains of oncogenes and losses of tumor-suppressor genes can both cause tumorigenic changes in a cell.

For deletion of a given diploid locus, there are only two options: deletion of one or two copies, also called hemi- and homozygous losses. Another nomenclature for deletions is loss of heterozygosity (LOH), which refers to a continuous stretch of DNA where all SNPs (single nucleotide polymorphisms)

appear homozygous. When such loss of one chromatid is compensated with a duplication of the sister chromatid, there is no alteration in the overall copy number. This phenomenon is termed copy-neutral loss of heterozygosity (cnLOH).

For gains, there can be a range of additional copies, which makes classification into exact copy number categories more complicated. In classical cytology terms, there is a division into gains and amplifications. Gains refer to one or a few additional copies, and can be as large as entire chromosomes. Amplifications are invariably focal and contain a large number of extra copies, up to over a hundred (Krijgsman et al., 2014).

Ploidy, cellularity, and heterogeneity

CNAs discussed in the previous section affect individual chromosomes, or parts thereof. In addition to aneuploidy, an abnormal number of chromosomes, cancers can also exhibit an abnormal number of complete sets of chromosomes. While normal somatic human cells are diploid and carry two copies of all autosomal chromosomes, some cancers have only one set, while others have multiple. Haploid (or monoploid) tumors have lost one set of chromosomes (Corver et al., 2014). This can be an intermediate step that precedes global cnLOH. Polyploid tumors, on the other hand, have gained additional copies of all chromosomes. This leads to the possibility of additional copy number levels that can be observed in measurement data. While a diploid cell can only undergo losses of one or both alleles, a tetraploid sample, for example, could exhibit four discrete levels of losses. This is similarly true for gains, and makes it progressively more challenging to identify the exact copy number behind a gain.

As described above, DNA copy numbers are fundamentally discrete integers. However, when measuring copy numbers of entire cancer samples with genome-wide techniques, these discrete integer values are usually obscured by admixed normal cells and intratumoral heterogeneity.

Cancer samples obtained from routine clinical care (biopsies or resections) rarely consist exclusively of tumor cells. Usually, they also contain a varying proportion of normal cells. Since a CNA is present only in the cancer cell population, its signal gets dampened in proportion to the percentage of infiltrating normal cells. This impacts its detection (Krijgsman et al., 2013). Since the proportion of cancer cells, the cellularity, varies, the same detection limits do not directly apply between samples. If the cellularities of individual samples are known, the dilutive effect can be corrected. However, accurate estimation of cellularities under a microscope can be challenging. Due to the aberrant size and shape of cancer cells, there is great inter-observer variability when pathologists judge cellularity. This is in contrast to procedures in clinical genetics and in the study of CNVs, where uniform cell populations can generally be assumed. Hence, while algorithms built primarily for CNVs or CNAs both measure the same phenomenon, namely DNA copy number, they often make different assumptions, and their performance can be suboptimal if used in the other setting.

In addition to infiltration with normal cells, the tumor itself can also be non-clonal (Burrell et al., 2013). Various distinct cell populations can be present either in separate parts of the tumor, or diffused together (Allison and Sledge, 2014; van Thuijl et al., 2014; Jamal-Hanjani et al., 2017). This results in a similar dampening of signals, as dictated by the proportion of cells carrying a particular aberration. It also introduces a mixture of copy number levels in the measurement data.

Review of literature for data analysis of CNAs

There are multiple techniques that can be used to detect alterations in DNA copy number, and the discipline is called cytogenetics (Hastings et al., 2016). Cytogenetic techniques can be divided in techniques that measure one or few chromosomal locations, and techniques that allow a genome-wide measurement. Modern genome-wide techniques produce vast amounts of data, and require sophisticated computational and statistical methods for their analysis and interpretation. This interdisciplinary field is known as bioinformatics.

This chapter (Chapter 1) reviews methods developed for the analysis of chromosomal copy number aberrations in cancer. It discusses the analysis of data obtained with two genome-wide high-throughput techniques, microarrays and next-generation sequencing (NGS). The focus is on experiments where the aim is to analyze larger series of patient samples in order to interpret and characterize the significance of the observed CNAs (as opposed to for example diagnostics where the aim is to detect the presence of absence of specific CNAs in individual patient samples).

The purpose of a study with larger series of patients can be for example to identify associations with clinical variables, to discover subtypes present in the data set, or to describe recurrent aberration patterns in a previously uncharacterized cancer type. In the following sections, analytical methods to achieve these goals that involve interpretation of CNAs in the context of a larger set of samples are referred to as *downstream analyses*.

Prior to proceeding to the *downstream analyses* to interpret the significance of CNAs in a data set, there are some preliminary analytical steps that need to be taken. There is no clear unambiguous consensus of which steps exactly need to be included, but in this text, they are collectively known as *copy number analysis*. This is not to imply that there is necessarily a clear separating boundary between the two, and that every type of analysis would clearly belong to one group or the other. The distinction is made solely on the conceptual level in order to facilitate discussion of the analytical workflow, and analysis of data originating from two different laboratory techniques. Whereas *downstream analyses* include standard statistical testing and learning procedures, *copy number analysis* is more dependent on the exact lab technique used to obtain raw data. Figure 1.1 shows an analytical workflow that starts from raw data and ends in statistical tests and learning methods. Intermediate steps will be added in the following sections.

In the next section, microarrays and their *copy number analysis* is discussed first. This introduces concepts and provides context that is useful for the subsequent review of different approaches for *copy number analysis* from NGS data. Finally, *downstream analyses* are discussed.

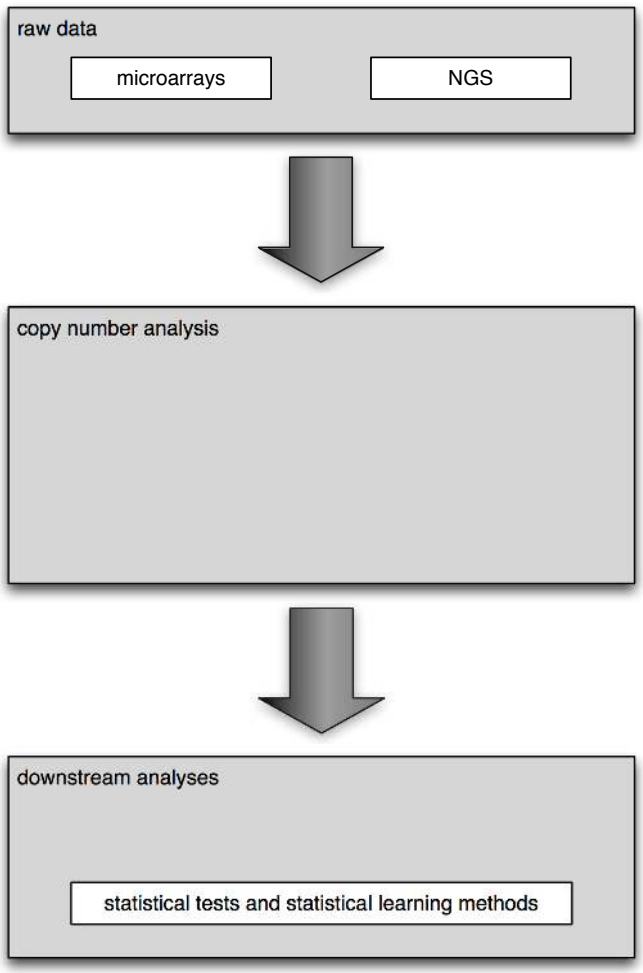

Figure 1.1: Analytical workflow. Starting from raw microarray or NGS data, the analytical workflow proceeds first through platform-dependent *copy number analysis* and ends in application of standard statistical tests and statistical learning methods. Intermediate steps are described in the following sections, and an updated version of the diagram presented at the end of this chapter.

Microarrays for genome-wide CNA detection

Array laboratory process

DNA microarrays were developed at the end of the previous century and consist of a set of DNA molecules with specific base pair sequences fixed on a solid medium. These array elements bind to molecules with the complementary nucleotide sequence, which can be either DNA or RNA.

Array comparative genomic hybridization (CGH) (Pinkel et al., 1998; Pollack et al., 1999) uses microarrays to measure DNA copy number. Its principle is similar to traditional chromosomal CGH (Kallioniemi et al., 1992), which uses metaphase spreads fixed on a glass slide. The elements on the array can be either BAC (bacterial artificial chromosome) clones, cDNA molecules, or synthetic oligonucleotides (Ylstra et al., 2006).

Briefly, the CGH array laboratory process is as follows. DNA isolated from normal and tumor samples are labeled with different fluorescent dyes. They are then allowed to simultaneously hybridize to the array, competing for the array elements. After unbound DNA is washed away, dye intensities are measured with a scanner. Finally, image analysis software is used to quantify the intensities of normal and tumor DNA.

In addition to CGH arrays, genotyping arrays are used for copy number detection. These were originally designed to detect SNPs with alternative short probes for the major and minor alleles. Compared to the exactly matching and longer oligonucleotides on CGH arrays, their signals for copy number detection were somewhat noisier. Therefore, newer generations of genotyping arrays, such as the Affymetrix Genome-Wide Human SNP Array 6.0, contain dedicated copy number probes in addition to the SNP probes. Genotyping arrays are typically hybridized with the tumor sample only, unlike the comparative CGH arrays (Ylstra et al., 2006). For copy number analysis, a separate normal signal is typically used, often in the form of a larger reference data set, such such as the HapMap project (International HapMap Consortium, 2003). B-allele frequencies (BAFs) from the SNP probes on genotyping arrays allow detection of cnLOH, which is not possible with CGH arrays. But when working with FFPE material, CGH arrays have been found to be more robust and generally outperform genotyping arrays (Curtis et al., 2009; Krijgsman et al., 2012).

Array data and meta-data

The exact format in which the obtained intensities are stored varies between manufacturers. A typical structure is a tabular text file, which, in addition to the numerical intensities, also contains additional quality control measurements or flags, and possibly background intensity measurements. After filtering potentially unreliable data points, and possible background correction (Ritchie et al., 2007), microarray data for copy number experiments is usually stored as $\log_2$-transformed ratios between the test and reference sample intensities. The ratios are presented as a matrix, where rows correspond to the elements of the array, and columns represent the individual specimens. When using genotyping arrays, an additional matrix of BAFs can also be used.

In addition to the data itself ($\log_2$-ratios and possibly BAFs), meta-data is also needed. Each element on the array needs to be annotated to define what it is measuring. Ideally, CGH arrays are annotated with the chromosome name and base pair location of each arrayed element. These can be provided by the array manufacturer, or obtained from the DNA sequences of the array elements with a tool such as BLAT (Kent, 2002). The elements can also be annotated with gene names, but while this works well for arrays that target only genes, such as cDNA arrays, copy number arrays usually contain elements that target genomic sequences lo-

cated between genes.

Publication of microarray results in a scientific journal usually requires that the original data is made publicly available in a database such as GEO (Gene Expression Omnibus; Edgar et al., 2002; Barrett et al., 2013) or ArrayExpress (Kolesnikov et al., 2015). Only by making the raw data and analytical procedures public, can others evaluate the reproducibility of final results (Ioannidis et al., 2009). In addition to the raw data and meta-data characterizing the array elements, meta-data describing the experimental conditions should also be included. These requirements are known as MIAME, Minimum Information About a Microarray Experiment (Brazma et al., 2001). Chapter 2 (Scheinin et al., 2008) will describe a public and MIAME-compliant database that is focused on CNAs and cancer.

Copy number analysis of microarray data

Raw microarray data needs to go through *copy number analysis*, which is broken down into discrete steps below, before it is ready for *downstream analyses*. *Preprocessing* is performed to account for technical artifacts, such as dye biases and possible missing values. But after *preprocessing*, there are differing strategies on whether data is ready for *downstream analyses* or if other subsequent steps are still required. These possible intermediate steps include *segmentation* (which refers to the identification of stretches of consecutive array elements that most likely share the same copy number and are separated by breakpoints) and *calling* (assignment of discrete copy number levels to the segments). Figure 1.2 outlines these steps, and the following sections describe them in more detail and discuss their benefits (van de Wiel et al., 2011).

Preprocessing of microarray data

In this text, the term *preprocessing* is used to refer to the steps that are needed to prepare the $\log_2$-ratios of individual samples for *segmentation* and *calling* algorithms. It can include steps such as filtering, imputation, normalization, and smoothing.

Filtering refers to unreliable values that should be removed. These can include data points flagged by array image analysis software due to various reasons, such as being saturated or non-uniform, or array elements for which it was later learned that their sequences bind to multiple positions in the genome. When this leads to missing values for some of the arrays in a data set, imputation can be used, with for example the k-nearest neighbors algorithm (Troyanskaya et al., 2001).

Normalization is used to eliminate various sources of bias, such as differences in the labeling efficiencies of the dyes. There are methods for both intra- and inter-array normalization, but the latter is usually not used in the analysis of CNAs due to natural variation in signal intensities between samples caused by varying cellularity, tumor heterogeneity, and ploidy. Normalization methods usually center the ($\log_2$-transformed) data around zero. The baseline at zero is interpreted as "normal" copy number, with gains and losses defined as relative to the baseline. In the case of polyploid or haploid samples, the copy number of the most common level naturally differs from the truly normal state of two copies for a diploid species. When using genotyping arrays instead of CGH, the BAFs can possibly be used to detect exact copy numbers, and to define "normal" as exactly two copies. Alternatively, an absolute copy number for a single chromosomal location can also be obtained with a technique such as FISH (fluorescent in-situ hybridization).

For identification of the baseline level, use of the mode would be ideal, but can potentially lead to ambiguous situations for computation algorithms in case of multi-modal distributions. Use of the median is there-

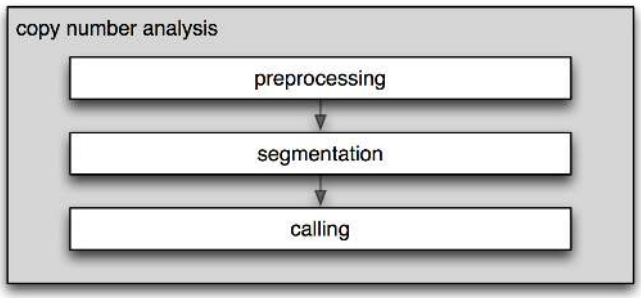

Figure 1.2: *Copy number analysis* of an array CGH sample. Three separate steps can be identified in the process. These include *preprocessing*, *segmentation*, and *calling*. Not all three are necessarily performed before proceeding to *downstream analyses*, as will be discussed later.

fore more common. More sophisticated normalization methods have also been published based on lowess (Staaf et al., 2007), ridge regression (Chen et al., 2008), better correction for spatial bias (Neuvial et al., 2006), or on a stepwise framework that independently targets intensity, spatial, plate, and background biases (Khojasteh et al., 2005).

If the proportion of cancer cells in the sample is known, some software packages, such as CGHcall (van de Wiel et al., 2007), allow the data to be adjusted for the dilutive effect of normal cells. This can be beneficial when samples in the same data set vary in terms of their cellularities.

Figure 1.3 shows an example array CGH profile of a low-grade glioma (LGG) sample that has been *preprocessed* by filtering outliers and median-normalized.

De-waving is an optional *preprocessing* step to correct an occasional artifact that can be observed as a wavy pattern, when copy number data is ordered and plotted according to their chromosomal position. It is related at least to GC content (the proportion of guanine and cytosine bases), but might also be affected by other factors (Marioni et al., 2007), including cell cycle and dye bias. During *preprocessing*, waves can optionally be corrected either by regression on GC content (Diskin et al., 2008), or using a calibration data set (van de Wiel et al., 2009). The benefit of the regression approach is that it does not require calibration data, whereas the calibration data approach can also correct for other sources of bias besides GC content.

Smoothing is another optional step, which reduces technical variation by utilizing spatial information of array elements along the chromosomes. The simplest approach is to use a moving window around the array element of interest and calculate the mean or median (CGH-Explorer, Lingjaerde et al., 2005; CLAC, Wang et al., 2005; ChARM, Myers et al., 2004). Other smoothing methods include quantile smoothing (quantreg, Eilers and de Menezes, 2005), adaptive weights smoothing (GLAD, Hupé et al., 2004, and wavelet smoothing (Hsu et al., 2005). Smoothing helps reduce noise, but also decreases sensitivity and increases risk of missing focal aberrations. Since smoothing dampens signals on both sides of copy number breakpoints, it can make it harder for *segmentation* algorithms to perform well and detect the exact breakpoint locations.

Smoothing can sometimes also refer to procedures that target only individual outliers. *Segmentation* methods that utilize *segment* means, such as the circular binary segmentation (CBS) algorithm, can be sensi-

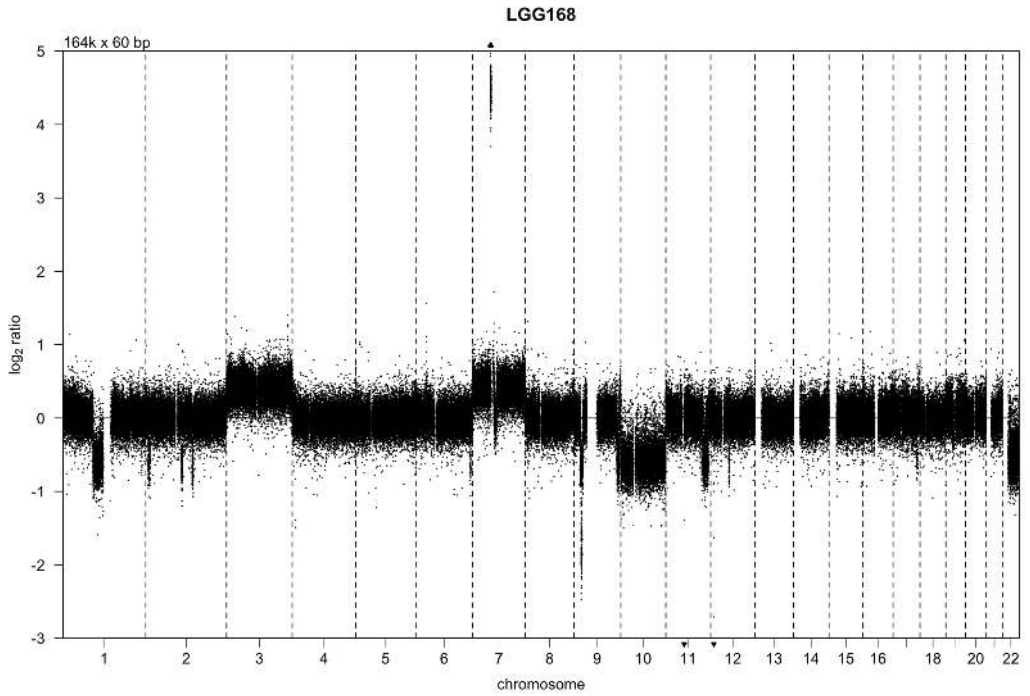

Figure 1.3: A *preprocessed* array CGH copy number profile. Array elements are ordered along the x-axis by their genomic positions, and y-axis shows median-normalized $\log_2$-ratios between an LGG sample and a normal reference sample. Small triangles at the top and bottom edges represent data points that fall outside the plot area. Upper left corner show the number of array elements (approximately 164,000) and their length (60 base pairs). This profile has not been de-waved since waviness is minimal. The sample shown here is from the LGG study that is presented in Chapter 5 (van Thuijl et al., 2014). Although that study uses NGS for CNA detection, CGH arrays were performed for some of the samples for quality control purposes.

tive to local outliers. Its implementation in DNAcopy (Venkatraman and Olshen, 2007) therefore first detects singleton outliers and shrinks their values towards their neighbors.

Segmentation and calling of microarray data

Algorithms designed to identify the locations of possible CNAs can be used after the $\log_2$-ratios have been *preprocessed*. This task can be further divided into two discrete steps: *segmentation* and *calling*. Although some algorithms perform both simultaneously, such as those based on hidden Markov models (HMMs; Eddy, 2004), they can be conceptually distinguished from each other, and are therefore discussed separately.

Segmentation refers to the identification of stretches of consecutive array elements ordered by chromosomal position that: 1) most likely share the same copy number, and 2) are separated by breakpoints. Their identification reduces noise and improves sensitivity and specificity (Willenbrock and Fridlyand, 2005). Another potential benefit is better detection of the correct baseline, the level that represents normal copy number. If few segments coincide with the level at zero, the baseline can be adjusted to fit the most common segment level. This is referred to as post-segmentation normalization.

Multiple *segmentation* algorithms have been developed. The most widely used is CBS (Olshen et al., 2004). Other examples include aCGH-Smooth (Jong et al., 2004), CGHseg (Picard et al., 2005), and mBPCR (Rancoita et al., 2009). These methods follow an approach with a local focus. Their aim is to identify breakpoints between neighboring segments, and these segments can be described with their means or medians. An alternative is a global, model-based approach, that aims to identify a set of copy number levels that fits all segments across the entire genome. One example of this approach is BioHMM (Marioni et al., 2006).

The way segmented data is stored varies between methods. Since the breakpoints between segments vary from one sample to another, it is not possible to simply use a similar matrix as with *preprocessed* data, but now with rows representing segments (except in the case of a single sample). Instead, the *preprocessed* data matrix is typically combined with vectors of segment boundaries for each sample. This can be further converted to a matrix of the original dimensions by replacing the $\log_2$-ratio of each array element with the corresponding segment mean, which makes the data straightforward to use in *downstream analyses*.

Calling represents the assignment of a discrete copy number state to each *segment*. The term *interpretation* is also sometimes used for this purpose. The simplest approach is to use cutoffs to define when *segmented* $\log_2$-ratios are too large or small to be considered "normal", and instead represent "gains" and "losses", respectively. Correct cutoff values, however, are affected by cellularity, possible heterogeneity, and tumor ploidy. These are often not precisely known and also vary between samples.

Due to limitations of cutoff-based *calling*, more sophisticated *calling* algorithms have been developed, based on various types of statistical models. Wang et al. (2005) were the first to publish such a method with the introduction of the CLAC algorithm. Many other approaches have been published since, and comparisons and reviews have been published by Lai et al. (2005) and Wang (2009). Compared to simple cutoffs, algorithms based on statistical models generally produce more reliable results with less false positives.

Depending on the algorithm, a *call* can refer to an *absolute* number of copies, such as "1" for a loss of one allele and "4" for a gain of two extra copies. But more commonly, *calls relative* to the most common, "normal", level are used. The minimum is three levels: "loss", "normal", and "gain", and further characterization is possible by separately including "ho-

mozygous deletions" and "amplifications".

Calling helps the interpretation of results from copy number experiments, as it describes the findings in a biologically more intuitive way compared to $\log_2$-ratios. It also makes it more straightforward to make comparisons between arrays, platforms, and experiments. Furthermore, it conceptually validates verification of results with another techniques, such as FISH (Snijders et al., 2001; van den IJssel et al., 2005), that verifies the copy number *call*, not a $\log_2$-ratio.

While *calls* are easier to interpret, they also result in loss of information, as continuous $\log_2$-ratios are replaced with discrete *calls*. Due to issues with cellularity, heterogeneity, and polyploidy, this information can be valuable. Multiple *calling* algorithms are based on mixture models, including CGHmix (Broët and Richardson, 2006), CGHclassify (Engler et al., 2006), and CGHcall (van de Wiel et al., 2007). In addition to discrete *hard calls*, these methods also return associated probabilities. They capture the related uncertainties caused by, for example, tumor heterogeneity. Use of *call probabilities*, or *soft calls* as they are sometimes referred to, can therefore combine the advantages and interpretability of *hard calls* without suffering from the associated information loss. An example of *segmented* and *called* copy number profile is shown in Figure 1.4 for chromosomes 7 and 9 of the same LGG sample as in Figure 1.3.

Unambiguous estimation of *absolute calls* from $\log_2$-ratios alone may be impossible due to issues with cellularity, heterogeneity, and ploidy. The use of genotyping arrays provides additional information in the from of BAFs, which help estimation of *absolute* copy numbers (Attiyeh et al., 2009; Rancoita et al., 2010; Ortiz-Estevez et al., 2012), and also aid detection of breakpoint locations during segmentation (Olshen et al., 2011).

Most *segmentation* and *calling* algorithms have been developed to handle data measured with a single microarray platform at a time, but there are also methods developed to combine multiple platforms, such as MSCN (Bengtsson et al., 2009) and MPCBS (Zhang et al., 2010). Chapter 2 will also describe one approach to deal with multiple array platforms, by focusing on genes instead of array elements.

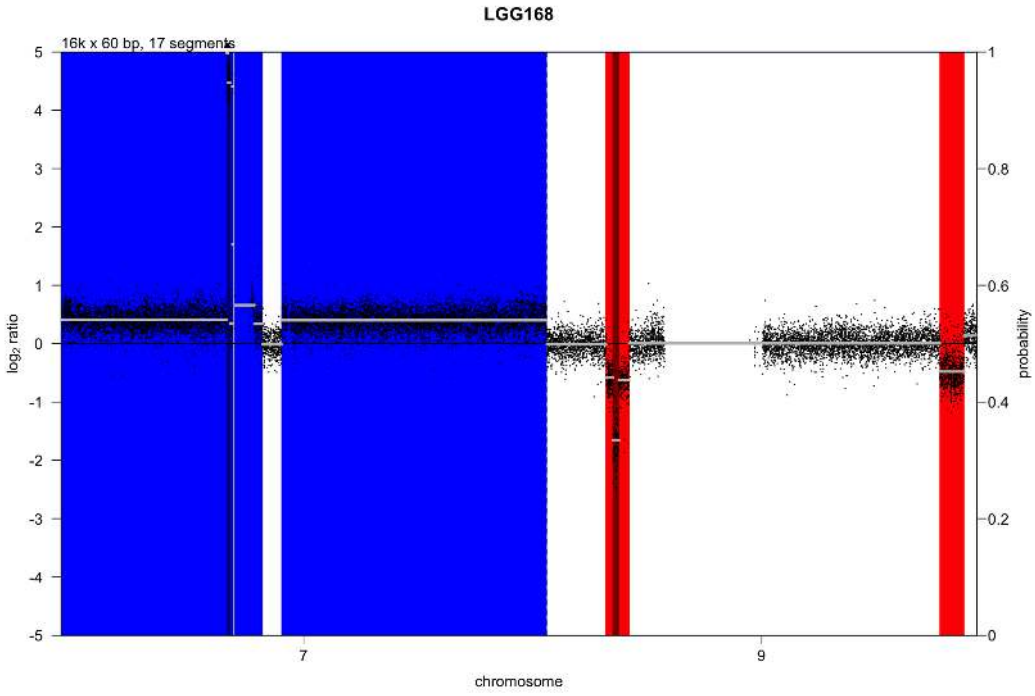

Figure 1.4: *Segmented* and *called* array CGH copy number profile. Chromosomes 7 and 9 of the same low-grade glioma sample as in Figure 1.3. Gray horizontal bars represent *segment* means, and the total number of *segments* detected is shown in the upper left corner. *Calling* results are shown with colored bars. Single-copy "losses" are shown in bright red, and "homozygous deletions" in dark red. These bars start from the bottom of the plot area and the associated call probabilities can be read from the right-hand scale. Similarly, "gains" are shown in bright blue and "amplifications" in dark blue, with bars that start from the top of the plot area and probabilities of 1 - the value of the right-hand scale. As all of the bars essentially cover the entire plot area, the *call probabilities* are very close to one and contain little uncertainty.

Next-generation sequencing for CNA detection

Sequencing laboratory process

The development of next-generation sequencing (NGS), or high-throughput sequencing, methods allows the sequences of DNA and RNA molecules to be deciphered much more quickly and cheaply (Koboldt et al., 2010). These methods have revolutionized many aspects of bio(medical) research, including cancer genomics and personalized medicine, with many new predictive and prognostic markers discovered (Meyerson et al., 2010).

NGS methods include technologies from a number of companies, with Illumina, Roche 454, and SOLiD as the most popular systems so far. Details of their respective procedures do vary from each other, but the common factor is that each sequences fragments of DNA in a massively parallel manner. From the perspective of data analysis, the practical differences are limited to the number of sequence reads obtained per run and the lengths of the reads. To explain the basic principles of NGS, Illumina sequencing is used here as an example, as it is the most widely used technology, and the original publications included as Chapters 4 and 5 of this dissertation are also based on Illumina data.

Briefly, isolated DNA is sheared to smaller fragments, and adaptor molecules ligated to each fragment. These adaptors allow the fragments to bind to the surface of a solid medium, the flow cell. PCR (polymerase chain reaction; Saiki et al., 1988) is then performed to multiply each individual molecule into a cluster of DNA strands with the same sequence to enhance detection. The flow cell is filled with DNA polymerase and fluorescently labeled nucleotides that have been modified to include a terminator, so that only one base is incorporated at a time. A scanner scans the flow cell separately for each of the four dyes, and the fluorescent signals indicate which base has been incorporated into which cluster. Once the terminators and fluorescent dyes have been detached from the clusters, the process is repeated. After a predetermined number of cycles, such as 50 or 100, have been performed, the fragments can optionally be turned over, and also their other ends sequenced in a similar fashion to obtain paired-end data.

Sequencing experiments can be divided into several categories. Whole-genome sequencing (WGS) is used to study entire genomes, either as *de novo* experiments to assemble a novel reference genome sequence, or more commonly as re-sequencing experiments to identify how particular individuals differ from an existing reference genome. Instead of sequencing the entire genome, targeted experiments can also be performed that focus only on specific areas. Before the prepared DNA (or RNA) library is placed on the flow cell, a subset of the fragments can be captured and selected for sequencing. The targeted subset can be for example the exome (whole-exome sequencing, WES), or a specific set of genes, such as the kinome (Majewski et al., 2013) or a panel of cancer genes (Sie et al., 2014). Since only a subset of the genome is sequenced (about 1 % for WES), this reduces costs and can therefore allow for larger sample sizes.

A single Illumina instrument run produces millions or billions of reads, depending on the instrument model. This total capacity can be used for a single DNA sample, but is often divided between multiple samples with a technique referred to as multiplexing. For multiplexing, each sample is assigned a unique barcode sequence that is included in its adaptor molecules. These sequences allow data originating from different samples to be distinguished. There is no *a priori* exact amount of data that can be generated per sample. Depending on the goal of the experiment, it can be flexibly controlled through the number of samples that are multiplexed together on each sequencing lane of a single instrument run.

An important concept related to instru-

ment throughput is the required amount of sequencing (or read) coverage, which refers to the number of times a single base position has been sequenced. The higher the coverage, the higher is also the confidence with which variations from a reference genome can be detected. The required minimum threshold depend on the application, statistical approaches used, and the accuracy of the sequencing technology in question (Glenn, 2011; Wang et al., 2012). For Illumina, the current error rate is approximately 0.5 %, and sequence coverage of at least 10× or 30× is generally recommended. However, the observed coverage can vary greatly across different parts of the genome (or targeted areas of the genome), so average coverage exceeding 100× can be needed to adequately cover some areas (Ross et al., 2013).

The quality of isolated DNA also affects error rates and sequence coverage requirements. While DNA from archival FFPE material is generally compatible with sequencing protocols (Schweiger et al., 2009; Wood et al., 2010), deep sequencing might reveal positions in the genome that are particularly degraded. Also, the enrichment protocols for WES or other types of targeted experiments may show uneven performance with FFPE material (Hedegaard et al., 2014).

NGS data and meta-data

A character string of DNA sequence (of the same length) is produced for each cluster once the scanned images have been processed. Meta-data is also recorded for each base in the form of a quality score, which represents the reliability of the base call in question. After the sequence reads have been obtained, they can be trimmed to remove unwanted sequences, such as those of the adaptor molecules. Then, except for *de novo* experiments, the reads are usually aligned to a reference genome with an algorithm such as BWA (Li and Durbin, 2009) or Bowtie (Langmead et al., 2009; Langmead and Salzberg, 2012).

Public access to raw data is the only way to ensure transparent reproducibility of published scientific experiments (Ioannidis et al., 2009). Raw sequencing data can be submitted to publicly accessible databases such as The Sequence Read Archive (SRA; Kodama et al., 2012) or The European Nucleotide Archive (ENA; Leinonen et al., 2011). Public availability of raw data is usually required in order to publish results of microarray experiments in scientific journals, but with NGS this requirement is perhaps not (yet) as widely enforced. One factor that may contribute to the somewhat more relaxed standards could be the higher storage requirements. In addition, (deep) sequencing data can potentially be highly identifiable, and privacy of the test subjects need to also be taken into account.

Approaches for copy number analysis by NGS

A number of methods have been published that detect copy number alterations from NGS data (Teo et al., 2012; Liu et al., 2013; Zhao et al., 2013). Based on the underlying principles, they can be grouped in five general categories: 1) paired-end mapping, 2) split-read, 3) depth of coverage, 4) assembly-based, and 5) combinatorial methods (Alkan et al., 2011). Each approach has its set of strengths and limitations, and choice depends on research goals and financial resources. The following sections describe the basics of each category and their applicability for cancer and CNAs.

Comparisons of algorithms are often based on simulations. Simulated data makes it straightforward to measure sensitivity and specificity of computational algorithms, as the underlying truth is known. While simulators have been developed that reproduce known biases from sequence context and empirical platform-dependent errors (including ART, Huang et al., 2012; pIRS, Hu et al., 2012; GemSIM, McElroy et al., 2012; and Wessim, Kim et al., 2013a), there is no comprehensive cancer simulator that would capture all the characteristics and complexities

of tumor genomes. Therefore, instead of simulations, data generated from biological samples is often used, and the challenge becomes to define which CNAs are real and which are not. Array CGH can be used as a reference point, but limited resolution and lack of reproducibility between array platforms and processing algorithms makes it difficult to achieve an unambiguous gold standard (Pinto et al., 2011). Instead of benchmarked performance, choice of method therefore often relies on other factors, such as experimental design, and the algorithm's description, technical requirements and implementation details (Liu et al., 2013).

Paired-end mapping methods

The first methods to detect copy number alterations from NGS data were based on discordant read pairs. They are referred to as paired-end mapping (PEM) methods, or sometimes read-pair methods. As the name implies, PEM methods require paired-end data, *i.e.* both ends of the DNA fragments need to be sequenced.

Discordant pairs refer to two sequence reads from both ends of a single DNA molecule that do not align to the reference genome as expected. If the distance between the read pair differs from the general distribution of insert sizes, this could have been caused by an insertion or deletion of genetic material in-between. A read pair that aligns to two different chromosomes is an indication of a translocation. A third possibility is an incorrect orientation, indicating an inversion event. Korbel et al. (2007) were the first ones to describe a PEM-based method to detect copy number alterations, later released the PEMer software tool (Korbel et al., 2009). Other published algorithms include BreakDancer (Chen et al., 2009), and VariationHunter (Hormozdiari et al., 2009).

PEM methods are fairly accurate in the identification of breakpoint locations. But as they require copy number events to occur between the two sequenced ends of a DNA fragment, they also have significant limitations. For example, CNAs of entire chromosomes are not detected with PEM, since they do not result in any discordant read pairs. In general, the average insert size sets an upper boundary for the size of detectable insertions (Medvedev et al., 2009). Therefore, the primary application of PEM methods is not in somatic CNAs, but in the study of hereditary CNVs. In addition to alterations of copy number, they are also able to detect other types of structural variants (SVs), such as inversions (Sudmant et al., 2015).

Split-read methods

Split-read methods rely on breakpoints located within reads, whereas PEM methods identify cases where a breakpoint has occurred between the two reads from both ends of a DNA fragment. When a breakpoint is contained within a read, the read might not align to the reference genome, or might do so only partially. Such reads are therefore candidates for split-read methods. To identify the exact breakpoint location, short subsets from the beginning and end of the read are aligned independently. Each one is then grown until the exact breakpoint is found. As split-read methods require breakpoints to be within reads, they depend heavily on read length. They are therefore more suitable for sequencing technologies that produce longer reads, such as Roche 454 pyrosequencing. Split-read methods depend on unique read mapping, and reads that align equally well to multiple locations in the genome, *multi-reads*, are problematic.

The first published split-read method was Pindel (Ye et al., 2009). One challenge for split-read methods is the high computational effort required to align very short sequences. Pindel uses paired-end data facilitate this. When one read of a pair aligns while the other one does not, the location of the aligned read can be used to reduce the search space for the candidate read, and thus the computational overhead. AGE incorporates its own alignment tool that targets only predefined SV regions (Abyzov and Gerstein, 2011). Usu-

ally split-read methods utilize WGS data, but SLOPE was developed for targeted data and a limited number of genomic regions of interest (Abel et al., 2010).

Depth of coverage methods

The NGS approach that is conceptually most similar to CGH microarrays, is termed depth of coverage (DOC). The amount of read coverage, or depth, along the genome is measured by counting reads. These counts reflect the underlying DNA copy number, and can be thought of as analogous to fluorescent signal intensities of CGH microarrays. A conceptual difference compared to both PEM and split-read methods is that they detect breakpoints (at the boundaries of copy number alterations), while DOC detects the copy number alteration (between breakpoints). Although there are also methods based on density (such as SeqCBS; Shen and Zhang, 2012), DOC methods usually divide the genome into bins and count the number of reads in each bin. How the binning is defined varies between methods. The number of reads in each bin can be fixed to a specific value, and size of bins can be varied accordingly along the genome, such as with SegSeq (Chiang et al., 2009). More commonly, the size of bins is fixed across the genome. The bins can either overlap, such as with CNV-seq (Xie and Tammi, 2009), or be adjacent to each other. Some algorithms let the user choose the bin size, while others determine it automatically based on the amount of sequence data. When the aim is to analyze a data set consisting of multiple samples, it can facilitate *downstream analyses* if the same bin size is used for all samples. The underlying principle behind all DOC methods is that the obtained sequence coverage along the genome depends on the underlying copy number. However, there are also various biases that affect the observed coverage (Aird et al., 2011; Nakamura et al., 2011; Ross et al., 2013; Taub et al., 2010). Different authors have taken different approaches on how to handle them.

Due to its cost-efficiency, WES is an attractive alternative to WGS for re-sequencing experiments that aim to detect cancerous mutations within coding regions of the genome. Although some methods, such as Control-FREEC (Boeva et al., 2012), can be used with both WGS and WES data, many methods cannot handle WES data due to its discontinuity and biases introduced by the capture process. Dedicated tools have therefore been developed for WES data, such as ExomeCNV (Sathirapongsasuti et al., 2011), VarScan 2 (Koboldt et al., 2012), ABSOLUTE (Carter et al., 2012), and EXCAVATOR (Magi et al., 2013). Guo et al. (2013) have published a comparison of WES-based copy number tools and array CGH, although their emphasis is on CNVs instead of CNAs. An alternative approach to DOC-based copy number detection from WES (or other types of targeted) data was taken by Kuilman et al. (2015). They utilize off-target reads that align to the reference genome outside the targeted areas, and can thus be thought of to represent data from a low-coverage WGS experiment. Talevich et al. (2016) extended this approach to utilize both on- and off-target reads.

The workflow of DOC methods is covered in more detail in a subsequent section. This is because of its overall importance for the study of CNAs, and also because the upcoming Chapter 4 describes a DOC method, and Chapter 5 an application of this method to a set of LGG samples.

Assembly-based methods

Cancer genomes can undergo massive rearrangements, and the most flexible and comprehensive approach to detect them is to use assembly-based methods. Instead of relying on read alignment to an existing reference genome, these methods use overlapping reads to assemble contigs, continuous sequences constructed from smaller fragments (Staden, 1979). The contigs can then be compared to the reference genome.

Instead of assembling completely *de novo*,

a reference genome can be used as a guide, which can improve computation times and contig quality. This is referred to as "comparative genome assembly" (Pop et al., 2004). Assembly-based methods include Velvet (Zerbino and Birney, 2008), ABySS (Simpson et al., 2009), SOAPdenovo (Li et al., 2010), Cortex (Iqbal et al., 2012), and Magnolya (Nijkamp et al., 2012).

While assembly-based methods offer a potentially unbiased approach to discover novel variants, they have large computational demands and require high sequence coverage (40×) (Li et al., 2010). In practice, this makes them suitable only for the (thorough) analysis of individual or small number of cases, instead of larger data sets. Also, highly repetitive parts of the genome are challenging to assemble, and additional techniques that produce longer continuous sequences might be needed, if such areas need to be covered as well.

Combinatorial methods

There are also packages that are based on a combination of approaches presented above, most commonly PEM with another category. The most common combination is to use DOC to detect copy number alterations, and further improve breakpoint detection by incorporating PEM. These methods include SDDetect (Zeitouni et al., 2010), CNVer (Medvedev et al., 2010), Genome STRiP (Handsaker et al., 2011), GASVPro (Sindi et al., 2012), and inGAP-sv (Qi and Zhao, 2011). Another approach is the combination of PEM and split-read methods. However, these are primarily utilized to detect SVs, not CNAs. Such methods include NovelSeq (Hajirasouliha et al., 2010) and SVseq (Zhang and Wu, 2011). HYDRA combines a PEM-based method and local *de novo* assembly (Quinlan et al., 2010), and BreaKmer is a combination of split-read and local assembly (Abo et al., 2015).

Compared to PEM, split-read methods provide more accurate detection of breakpoint locations and therefore offer another complement for DOC. Such methods have been presented by Nord et al. (2011) and McKernan et al. (2009). Since PEM methods generally require WGS data, the combination of DOC and split-read methods are more suitable for target-enriched, such as WES, data.

Copy number analysis of DOC data

After briefly describing different approaches of copy number detection from NGS data, this section takes a more detailed look into the analysis of DOC data, for three reasons. First, two of the original publications included in this dissertation are based on DOC. Second, since DOC bears many conceptual similarities to the analysis of microarray data, some of what was covered on microarray analysis in the beginning of this chapter also be applied to DOC copy number data as well. And third, as will be discussed later, similar approaches for *downstream analyses* can be used for both microarray and DOC data.

The structure of bin-level DOC data is very similar to that of microarrays; a matrix with rows corresponding to bins along the genome and columns representing individual samples. The difference is that instead of fluorescent intensities, the matrix contains read counts. In general, its analysis also follows a similar path as with microarrays: *preprocessing* followed by *segmentation* and *calling* (Figure 1.2).

Preprocessing corrects for biases, such as GC content and mappability, and filters out problematic areas, such as centromeres, telomeres, and other repetitive sequences. *Segmentation* and *calling* identify breakpoints and estimate copy numbers. Available software packages vary in whether they contain all or part of these steps.

Preprocessing of DOC data

The most-characterized bias in observed sequencing coverage is caused by GC content (the proportion of guanine and cytosine bases), as first described by Bentley et al. (2008). Most DOC methods calculate the

GC content of each bin, and use the value to correct the read count for the bin (Yoon et al., 2009). Benjamini and Speed (2012) have shown that this is a good approximation in most cases, but can be further improved by separately calculating the GC content of each DNA fragment that has been sequenced. This requires paired-end data so that both end points, and therefore the entire sequence, of the fragment are known.

Another widely recognized source of bias is mappability, which refers to the uniqueness of sequences in the reference genome. It depends on the length of the sequence and the number of mismatches allowed (Whiteford et al., 2005). If $F_k(x)$ is the frequency at which the k-mer sequence at position x in the genome can be found in the entire reference genome (and its reverse complement), the mappability of this position is defined as $M_k(x) = \frac{1}{F_k(x)}$. Regions with low mappabilities contain highly repetitive sequences, which are problematic for all copy number detection methods. Excluding them from the analysis helps reduce false positives. Mappabilities can be calculated with programs such as GEM tools (Derrien et al., 2012). DOC methods typically use the average mappability of a bin for corrections and/or filtering.

The simplest approach to handle biases, is to require a separate normal reference sample, and to divide the read counts of the test sample with those of the reference. Since the bin GC contents and mappabilities are naturally the same for both samples, they could both be assumed to be similarly affected. However, as the GC content bias depends on the GC content of the molecules being sequenced, it is affected by the distribution of fragment lengths (Benjamini and Speed, 2012). If these vary between DNA libraries, the extent of the GC content bias will also be affected. Therefore, relying on a ratio between test and reference samples is an insufficient approach to handle this bias. Also, an approach such as the one taken by CNAnorm of performing a correction for GC content, but only after taking the ratio is suboptimal (Gusnanto et al., 2012). Ideally, GC content bias should be corrected separately for each sample, including possible reference samples.

Not all DOC methods require normal references samples. Two such algorithms are FREEC (Boeva et al., 2011) and readDepth (Miller et al., 2011). They both perform corrections for GC content and mappability, and also filter out low-mappability regions. Chapter 4 (Scheinin et al., 2014) describes the QDNAseq method, which also is also based on DOC and does not require a reference sample.

Segmentation and calling of DOC data

As in the previous sections, here *segmentation* and *calling* refer to identification of continuous stretches along chromosomes that most likely share the same copy number and are separated by breakpoints, and assignment of discrete copy number levels to these *segments*. Similar algorithmic approaches are used for these purposes as with microarrays.

Patchwork (Mayrhofer et al., 2013), ExomeCNV (Sathirapongsasuti et al., 2011), and VarScan 2 (Koboldt et al., 2012) all perform *segmentation* with the popular CBS algorithm, which was originally developed for the analysis of CGH microarrays (Olshen et al., 2004; Popova et al., 2009; Olshen et al., 2011). OncoSNP-SEQ (Yau, 2013) and HMMCOPY (Ha et al., 2012) are both based on HMMs, another popular analytical approach for CGH arrays. For *calling*, many methods are based on simple cutoffs, including BIC-seq (Xi et al., 2011), ExomeCNV (Sathirapongsasuti et al., 2011), FREEC (Boeva et al., 2011), readDepth (Miller et al., 2011), SegSeq (Chiang et al., 2009), and VarScan 2 (Koboldt et al., 2012). Due to challenges caused by cellularity, heterogeneity, and polyploidy, cutoffs often need to be defined on a case-by-base basis, and more sophisticated approaches based on statistical models could result in similar improvements in *calling* as with microarrays. One reason for the use of cutoffs could be that

many copy number methods for NGS are focused on detecting aberrations in individual samples, not for the concurrent analysis of larger data sets.

Like CGH microarrays, low-coverage DOC methods are unable to detect cnLOH. When sequence coverage is deep enough for variant calling, DOC data can be supplemented with variant allele frequencies. These aid in detection of cnLOH, and can also allow estimation of cellularity (for example Patchwork, Mayrhofer et al., 2013; CNAnorm, Gusnanto et al., 2012; OncoSNP-SEQ, Yau, 2013; HMMCOPY, Ha et al., 2012; and Control-FREEC, Boeva et al., 2012) and ploidy (Patchwork, CNAnorm, OncoSNP-SEQ, and HMMCOPY). As WES experiments have higher coverage, methods developed for WES data have typically been designed to incorporate allele frequencies in estimation of copy numbers.

For the analysis of cancer samples and detection of CNAs, it should be noted that some DOC copy number methods that have been developed specifically for germline CNVs might include assumptions and optimizations that make them unsuitable for somatic CNAs (Magi et al., 2012). For example, JointSLM (Magi et al., 2011), ERDS (Zhu et al., 2012), and CoNIFER (Krumm et al., 2012) reduce false positives by filtering out sharp, non-recurrent peaks, since they are thought to represent artifacts. If these methods are applied to tumor genomes, somatic CNAs might remain undetected.

Analysis of DOC data conceptually consists of the same steps as array CGH analysis: *preprocessing* followed by *segmentation* and *calling*. In microarray analysis, the implementation of these methods is often relatively modularized and they can be combined as desired. As an example, users can choose to perform wave corrections with either regression on GC content (Diskin et al., 2008) or using a calibration data set (van de Wiel et al., 2009), and free to choose their *segmentation* and *calling* methods of choice. However, many DOC methods are not designed to be modular in the same way. They perform their entire pipeline of *preprocessing*, *segmentation* and *calling*, and then produce one or more output files. Therefore, it can require extra effort to, for example, combine the *preprocessing* of one method with the *segmentation* and *calling* of another. The format and contents of these output files also vary between methods. For example, readDepth (Miller et al., 2011) returns only *segment*-level ratios and *calls* without full bin-level *preprocessed* data. It is therefore impossible to combine its *preprocessing* with other packages. FREEC (Boeva et al., 2011) on the other hand produces all three levels of *preprocessed*, *segmented* and *called* data, thus making it possible for users to combine it with other methods if they so wish.

The microarray community has benefited enormously from standardization and modularization. The popular Bioconductor suite (Gentleman et al., 2004; Huber et al., 2015) defined MIAME-compliant (Brazma et al., 2001) and extensible data structures for microarray data, which have been used by numerous packages. Due to the common data structures, these tools can be used in a modular fashion. Also, meta-packages that combine multiple tools together facilitating comparisons have been developed (Diaz-Uriarte, 2014) Due to the maturity of array technology, many analytical CNA packages are also relatively mature. They have accumulated years of usage from numerous labs with real experimental data. This has resulted, at least in the best-case scenarios, in improvements in efficiency and fine-tuning options, and also exposure and fixing of possible programming mistakes. *Segmentation* and *calling* methods that have been developed for arrays are therefore interesting candidates for the analysis of *preprocessed* DOC data as well.

Array methods usually assume the $\log_2$-ratios follow a Gaussian distribution, while read counts come from a Poisson distribution. However, most DOC methods do approximate the Poisson with Gaussian. Exceptions include readDepth (Miller et al., 2011),

which uses a negative-binomial, and BIC-seq (Xi et al., 2011), which is based on minimizing a modified BIC (Bayesian information criterion) and an approach where the distributional assumption cancels out of the equation. Therefore, use of the Gaussian distribution should not rule out use of array methods with DOC data. Furthermore, an appropriate transformation can be used. The Anscombe transform ($A : x \rightarrow 2\sqrt{x + \frac{3}{8}}$) transforms a Poisson distributed variable to one with an approximately standard Gaussian distribution (Anscombe, 1948).

Efforts towards shared data structures for DOC data would facilitate not only modular use of tools developed for DOC data, but also utilization of already established microarray methods. And since common data structures already exist for microarrays, they offer a potential candidate for developers of DOC-based methods as well. This would also facilitate *downstream analyses*.

Downstream analyses of CNAs

The term *downstream analyses* is used here to refer to procedures used not to detect CNAs of individual samples, but rather for the interpretation of CNAs in a series of cancer samples. This type of *downstream analyses* include standard statistical testing and learning procedures, such as hypothesis testing, clustering, and classification. This section discusses characteristics of copy number data, and corresponding analytical steps to be taken into account while applying these procedures. Most of the specific software packages mentioned in this section were originally developed for microarray data analysis, but can often be directly (or possibly with the Anscombe transformation) applied for DOC-based NGS data as well. Copy number alterations detected from NGS data with another type of method besides DOC can also easily be converted to a format suitable for these packages. The focus is on experiments that utilize copy number data alone. Integration of copy number with mRNA or miRNA expression (Louhimo et al., 2012), DNA methylation (Sokolova et al., 2016), or proteomic (Zhang et al., 2016) data is therefore not covered, although these integrative experiments can be crucial for identification of the driving cancer genes behind specific CNAs and thus better understanding of their significance.

Three topics that apply to all categories of *downstream analyses* are 1) the type of data, 2) the sample size, and 3) the dimensionality of the data, and these topics are briefly described first. 1) The type of data, refers to whether *preprocessed*, *segmented*, or *called* copy number data is used. There is no unambiguous consensus on which one should be used for *downstream analyses*, and there are examples of all three in published literature (van Wieringen et al., 2007).

2) Sample size refers to the number of biological samples (typically number of columns in the data matrix) in the data set. While the question of adequate sample size is relevant to all types of *downstream analyses*, it is most clearly defined in the context of hypothesis testing with statistical tests. Sample size calculations (or power calculations) can be used to define how many samples are needed to achieve a pre-determined level of statistical confidence. These calculations can be roughly divided to two groups: a) methods that ask the user to specify values for the required parameters (effect size, variance, desired significance level, and desired statistical power), and b) methods that estimate these parameters from a pilot data set. Chapter 3 (Scheinin et al., 2010) will present our efforts to develop a tool for sample size calculations in the setting of CNAs between two groups of cancer samples.

3) Dimensionality of the data refers to the number of genomic features (rows in the data matrix). These can be elements on a microarray or DOC sequence bins in an NGS experiment. The next section introduces an additional option, dimensionality-reduction through *regioning*, and discusses its potential benefits.

Regioning to reduce dimensionality

Individual copy number data points are dispersed along the chromosomes, and breakpoints can occur anywhere between (or within) these elements. This applies equally to data points that correspond to elements on a microarray or DOC bins from an NGS experiment, but also to data that has been transformed so that each known gene is represented as a separate data point. As the size of CNAs vary from entire chromosome arms to focal aberrations, an individual CNA can contain anything from thousands of data points to only one or two. For multivariate *downstream analyses*, such as clustering, this could have profound consequences as large CNAs can be given much more weight than smaller ones, even though the latter can be biologically just as important (Krijgsman et al., 2014).

Another challenge that is common with genomics data is known as "the curse of dimensionality". It refers to the fact that the number of biological samples included in a typical experiment is much smaller the number of genomic features measured per sample. This can lead to a massive multiple testing issue, which can be mitigated if the dimensionality can be reduced.

For copy number experiments, dimensionality can be reduced by *regioning*. That is, by identifying stretches of consecutive features along chromosomes that behave similarly across the entire data set, and collapsing them to single data points. Thus, *regions* can be thought of as a data-driven unit that aims to capture the underlying biological phenomena behind copy number alterations. The definition of similar behavior can be strict, or some loss of information can be allowed in the detection of breakpoint locations (van de Wiel and van Wieringen, 2007). An example of *regioning* results is presented in Figure 1.5, which shows a frequency plot of CNAs in 98 LGG patient samples.

Regioning in this sense requires the use of *called* data. Dimensionality can also be reduced with PCA (principal component analysis) on *preprocessed* data, and the resulting components though to represent supersets of genomic features. Interpretation of such data can be difficult, but the procedure can help to make visualizations in two or three dimensions. Similar factoring can also be performed with *called* data (Jöreskog and Moustaki, 2001).

Now that *copy number analysis* of microarray and DOC-based NGS data (Figure 1.2) and dimensionality reduction with *regioning* have been described, the workflow diagram presented in Figure 1.1 can be updated. The full analytical workflow from raw data into statistical tests and statistical learning methods is presented in Figure 1.6.

Identification of recurrent aberrations

The goal of some CNA studies is to identify the most frequent aberrations for a specific cancer type or subtype. If these CNAs are sufficiently small, it can be possible to pinpoint which of the genes within the aberration are the driving oncogenes or tumor-suppressor genes. This often requires integration of other types of information, such as which genes are expressed in the tissue or cell type in question under normal conditions, which genes located within gains or losses show over or under-expression, respectively, in the cancer samples, or functional assays to screen for tumorigenic activity of the gene products.

Rouveirol et al. (2006) first introduced the concept of recurrent minimal aberrations. If the *regioning* described in the previous section has been performed, the most common aberrations can be identified simply by comparing aberration frequencies of consecutive *regions*, and pinpointing *regions* with higher frequencies than their neighbors. A number of other approaches and software packages have also been described, including the popular GISTIC (Genomic Identification of Significant Targets in Cancer; Beroukhim et al., 2007) algorithm. The packages vary in their required input (*calls*, or *preprocessed* or *segmented* $\log_2$-ratios). Use of *called* data has the advantage of reduced noise, but may lead to lower sensitivity due to the loss of information compared to $\log_2$-ratios (Shah, 2008). On the other hand, use of $\log_2$-ratios can suffer from differences in magnitude between samples, caused by differences in cellularity, and possible polyploidy and heterogeneity. These issues can be reduced with the use of *call probabilities (soft calls)* that combine the noise-reducing effect of *calls* without losing information on the certainty of the aberration (Rueda and Diaz-Uriarte, 2009).

Statistical tests for association with clinical data

Many CNA studies aim to identify associations between specific aberrations and clinical variables. To detect CNAs associated with a clinical outcome, such as tumor progression, relapse, treatment success, or survival, stan-

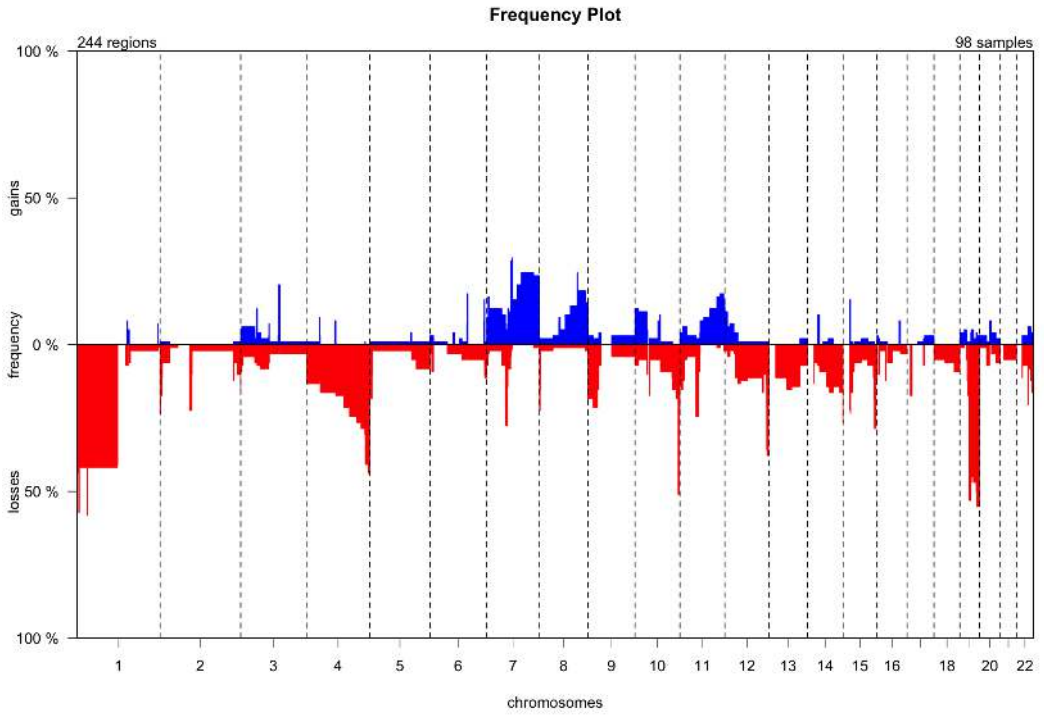

Figure 1.5: Frequency plot of CNAs. Frequencies of gains (shown with blue bars) and losses (red) of the 98 NGS samples of the LGG study (Chapter 5; van Thuijl et al., 2014). After *copy number analysis (preprocessing*, *segmentation*, and *calling)*, dimensionality has been reduced with *regioning*, resulting in a total of 244 *regions*.

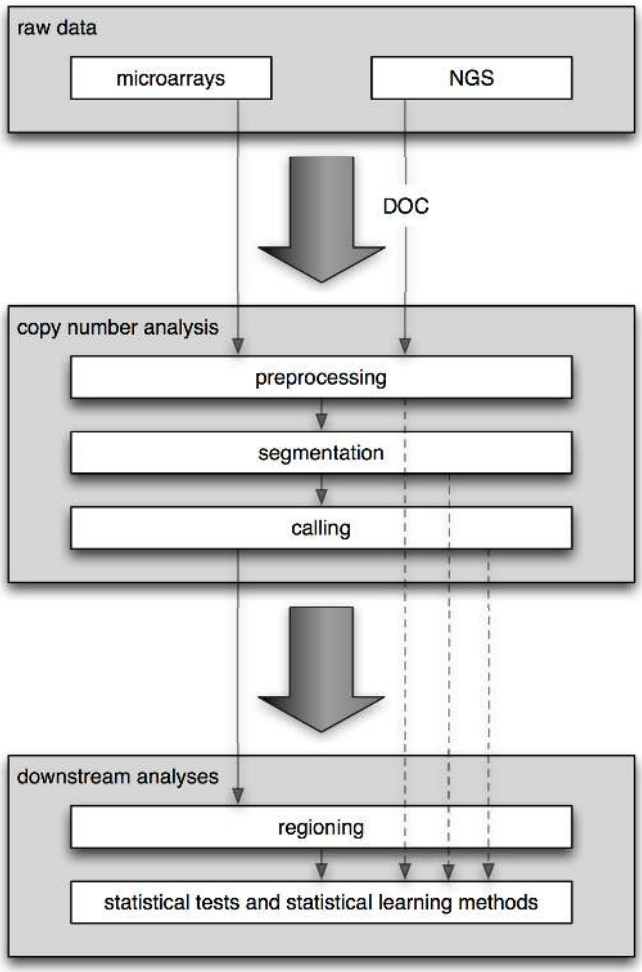

Figure 1.6: Full analytical workflow of CNAs. Starting from raw microarray or NGS data, the analytical workflow proceeds first through platform-dependent *copy number analysis*, which for microarrays and DOC-type NGS can be separated into *preprocessing*, *segmentation*, and *calling*. As a preliminary step in *downstream analyses*, dimensionality-reduction with *regioning* helps with multiple testing issues and better captures the underlying biology behind CNAs of various sizes. As there is no clear consensus on whether all included steps are necessary, the dashed arrows depict alternative workflows. Other approaches for NGS besides DOC are not included in the diagram.

dard statistical tests are used. Such tests can be performed on $\log_2$-ratios *(preprocessed* or *segmented)* or *called* data (per original genomic feature or per *region*).

Methods such as the standard Student's t-test assume Gaussian distributions. This assumption is not valid for $\log_2$-ratios of copy number data, because of the discrete nature of the underlying biology. Non-parametric, rank-based tests, such as Wilcoxon–Mann–Whitney test (Mann and Whitney, 1947), are therefore more suitable. But since they have been designed to have power for shift rather than multi-modality, tailored non-parametric tests could be more optimal (van Wieringen et al., 2008a).

When working with copy number *calls*, one can use common non-parametric tests, such as the χ^2 (chi squared) test for group comparisons, or the log-rank test for survival. These tests can be used for original genomic features or *regions*. As *regioning* reduces dimensionality, the multiple testing corrections are not as severe, which improves sensitivity (van de Wiel et al., 2005; van de Wiel and van Wieringen, 2007). Chapter 5 will cover an example of the log-rank test with *regions* for the association between CNAs and survival of LGG patients.

The use of discrete *hard calls* fails to take into account their possibly uncertainty (caused for example by the varying proportion of cells that contain a specific aberration), and therefore loses information compared to the *log*$_2$-level data. Higher power can be achieved with the inclusion of the *call probabilities (soft calls)* (van de Wiel et al., 2007; González et al., 2009).

Clustering for subtype discovery

Another common goal of copy number studies in cancer research is data-driven subtype discovery. Cluster analysis, or clustering, is an unsupervised learning method with the aim to group samples so that objects within a group (or cluster) are more similar to each other than to those in other groups. Various metrics can be used to measure this similarity (van Wieringen et al., 2008b).

As cancer is a heterogeneous group of diseases, it is well suited for cluster analysis. Cancers of a given tissue or organ typically consist of multiple different subtypes. Analysis of CNA patterns has been shown to be successful in identifying these subtypes (Chin et al., 2007b; Smeets et al., 2011). Similarly, meta-analyses of various cancer types has shown that cancers with similar etiology, cell-of-origin, or topographical location cluster together based on their CNA profiles (Myllykangas et al., 2006; Jong et al., 2007). Clustering of CNA data was initially performed with $\log_2$-ratios, first with *preprocessed* (Wilhelm et al., 2002) and then with *segmented* data (Jong et al., 2007). Tailored clustering methods for copy number data have since then also been developed. These include k-means (Liu et al., 2006) and hierarchical (van Wieringen et al., 2008b) clustering approaches. Both present distance measures that have been developed to deal with discrete data.

If the distance metrics are calculated from all original genomic features, larger aberrations are given considerably more weight than smaller ones. However, as the biological consequences of a small focal aberration can be as important as those of a whole-chromosome arm (Krijgsman et al., 2014), dimensionality reduction through *regioning* can lead to alternate, presumably improved, clustering (van Wieringen et al., 2008b; Liu et al., 2007). Chapter 5 (van Thuijl et al., 2014) will present examples of clustering for LGG patients based on copy number *regions* and *soft calls*.

In addition to the distance-based methods discussed above, model-based approaches have also been used, for example using HMMs (Shah et al., 2009). A comparison of clustering methods for copy number data has been published by Brito et al. (2013). An alternative approach for subgroup discovery is to use PCA (or another similar dimensionality reduction technique), and to identify clusters by visual inspection of low-dimensional plots

(Somiari et al., 2004; Unger et al., 2008). The downside is harder interpretability of differences between subtypes, because the resulting components ignore the order and location of the original features along the genome, and thus do not represent a entity with a biological interpretation.

Aims of this dissertation

The aim of my PhD was to develop bioinformatic solutions for copy number analysis in cancer. Broken down to separate projects that are covered by the original articles included in this dissertation as upcoming chapters, the aims were as follows:

1. To develop a database solution for storage of raw and processed copy number data from cancer samples. A solution should support queries based on clinical variables and specific copy number alterations. It should also be MIAME-compliant and publicly accessible. The developed solution, a database called CanGEM, is described in Chapter 2.
2. To develop a solution for sample size calculations for copy number experiments that compare two groups of cancer samples. A solution should allow the assessment of how many samples are required to satisfy statistical requirements. It should use pilot data to estimate the necessary model parameters. The CGHpower sample size tool is described in Chapter 3.
3. To develop a solution to detect copy number aberrations from cancer samples with NGS and the DOC approach. A solution should allow corrections for known systematic biases and filtering of problematic areas in the genome. It should also be compatible with archival material. The QDNAseq preprocessing method is described in Chapter 4, and an application is presented for a series of low-grade glioma patient samples in Chapter 5.

References

1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, and Abecasis GR, *et al.* (2015). A global reference for human genetic variation. *Nature*, **526**: 68–74.

Abel HJ, Duncavage EJ, Becker N, Armstrong JR, Magrini VJ, and Pfeifer JD (2010). SLOPE: a quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data. *Bioinformatics*, **26**: 2684–8.

Abo RP, Ducar M, Garcia EP, Thorner AR, Rojas-Rudilla V, Lin L, Sholl LM, Hahn WC, Meyerson M, Lindeman NI, Van Hummelen P, *et al.* (2015). BreaKmer: detection of structural variation in targeted massively parallel sequencing data using kmers. *Nucleic Acids Res*, **43**: e19.

Abyzov A and Gerstein M (2011). AGE: defining breakpoints of genomic structural variants at single-nucleotide resolution, through optimal alignments with gap excision. *Bioinformatics*, **27**: 595–603.

Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, and Gnirke A (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*, **12**: R18.

Albertson DG, Collins C, McCormick F, and Gray JW (2003). Chromosome aberrations in solid tumors. *Nat Genet*, **34**: 369–376.

Alkan C, Coe BP, and Eichler EE (2011). Genome structural variation discovery and genotyping. *Nat Rev Genet*, **12**: 363–76.

Allison KH and Sledge GW (2014). Heterogeneity and cancer. *Oncology (Williston Park)*, **28**: 772–778.

Anscombe FJ (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, **35**: 246–254.

Attiyeh EF, Diskin SJ, Attiyeh MA, Mossé YP, Hou C, Jackson EM, Kim C, Glessner J, Hakonarson H, Biegel JA, and Maris JM, *et al.* (2009). Genomic copy number determination in cancer cells from single nucleotide polymorphism microarrays based on quantitative genotyping corrected for aneuploidy. *Genome Res*, **19**: 276–83.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, *et al.* (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*, **41**: D991–D995.

Bell DW (2010). Our changing view of the genomic landscape of cancer. *J Pathol*, **220**: 231–43.

Bengtsson H, Ray A, Spellman P, and Speed TP (2009). A single-sample method for normalizing and combining full-resolution copy numbers from multiple platforms, labs and analysis methods. *Bioinformatics*, **25**: 861–7.

Benjamini Y and Speed TP (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res*, **40**: e72.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, *et al.* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**: 53–9.

Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S, Du J, *et al.* (2007). Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A*, **104**: 20007–12.

Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, *et al.* (2010). The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**: 899–905.

Blow N (2007). Tissue preparation: Tissue issues. *Nature*, **448**: 959–963.

Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, Janoueix-Lerosey I, Delattre O, and Barillot E (2012). Control-freec: a tool for assessing copy number and allelic content using next-generation sequencing data: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, **28**: 423–425.

Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, and Barillot E (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, **27**: 268–269.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, *et al.* (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, **29**: 365–371.

Brito I, Hupé P, Neuvial P, and Barillot E (2013). Stability-based comparison of class discovery methods for DNA copy number profiles. *PLoS One*, **8**: e81458.

Broët P and Richardson S (2006). Detection of gene copy number changes in CGH microarrays using a spatially correlated mixture model. *Bioinformatics*, **22**: 911–8.

Burrell RA, McGranahan N, Bartek J, and Swanton C (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, **501**: 338–345.

Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, Laird PW, Onofrio RC, Winckler W, Weir BA, Beroukhim R, *et al.* (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol*, **30**: 413–21.

Casparie M, Tiebosch ATMG, Burger G, Blauwgeers H, van de Pol A, van Krieken JHJM, and Meijer GA (2007). Pathology databanking and biobanking in The Netherlands, a central role for PALGA, the nationwide histopathology and cytopathology data network and archive. *Cell Oncol*, **29**: 19–24.

Chen HIH, Hsu FH, Jiang Y, Tsai MH, Yang PC, Meltzer PS, Chuang EY, and Chen Y (2008). A probe-density-based analysis method for array cgh data: simulation, normalization and centralization. *Bioinformatics*, **24**: 1749–56.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, *et al.* (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*, **6**: 677–81.

Chiang DY, Getz G, Jaffe DB, O'Kelly MJT, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, and Lander ES (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods*, **6**: 99–103.

Chin SF, Teschendorff AE, Marioni JC, Wang Y, Barbosa-Morais NL, Thorne NP, Costa JL, Pinder SE, van de Wiel MA, Green AR, Ellis IO, *et al.* (2007a). High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol*, **8**: R215.

Chin SF, Wang Y, Thorne NP, Teschendorff AE, Pinder SE, Vias M, Naderi A, Roberts I, Barbosa-Morais NL, Garcia MJ, Iyer NG, *et al.* (2007b). Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. *Oncogene*, **26**: 1959–1570.

Church DM, Lappalainen I, Sneddon TP, Hinton J, Maguire M, Lopez J, Garner J, Paschall J, DiCuccio M, Yaschenko E, Scherer SW, *et al.* (2010). Public data archives for genomic structural variation. *Nat Genet*, **42**: 813–814.

Corver WE, van Wezel T, Molenaar K, Schrumpf M, van den Akker B, van Eijk R, Ruano Neto D, Oosting J, and Morreau H (2014). Near-haploidization significantly associates with oncocytic adrenocortical, thyroid, and parathyroid tumors but not with mitochondrial DNA mutations. *Genes Chromosomes Cancer*, **53**: 833–844.

Cox C, Bignell G, Greenman C, Stabenau A, Warren W, Stephens P, Davies H, Watt S, Teague J, Edkins S, Birney E, *et al.* (2005). A survey of homozygous deletions in human cancer genomes. *Proc Natl Acad Sci U S A*, **102**: 4542–4547.

Croce CM (2009). Causes and consequences of microrna dysregulation in cancer. *Nat Rev Genet*, **10**: 704–14.

Curtis C, Lynch AG, Dunning MJ, Spiteri I, Marioni JC, Hadfield J, Chin SF, Brenton JD, Tavaré S, and Caldas C (2009). The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics*, **10**: 588.

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, *et al.* (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**: 346–352.

Dancey JE, Bedard PL, Onetto N, and Hudson TJ (2012). The genetic basis for cancer treatment decisions. *Cell*, **148**: 409–420.

Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, and Ribeca P (2012). Fast computation and applications of genome mappability. *PLoS One*, **7**: e30377.

Diaz-Uriarte R (2014). ADaCGH2: parallelized analysis of (big) CNA data. *Bioinformatics*, **30**: 1759–1761.

Diskin SJ, Li M, Hou C, Yang S, Glessner J, Hakonarson H, Bucan M, Maris JM, and Wang K (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res*, **36**: e126.

Dumanski JP, Rasi C, Lönn M, Davies H, Ingelsson M, Giedraitis V, Lannfelt L, Magnusson PKE, Lindgren CM, Morris AP, Cesarini D, *et al.* (2014). Smoking is associated with mosaic loss of chromosome Y. *Science*, **347**: 81–83.

Eddy SR (2004). What is a hidden Markov model? *Nat Biotechnol*, **22**: 1315–1316.

Edgar R, Domrachev M, and Lash AE (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**: 207–210.

Eilers PHC and de Menezes RX (2005). Quantile smoothing of array CGH data. *Bioinformatics*, **21**: 1146–1153.

Engler DA, Mohapatra G, Louis DN, and Betensky RA (2006). A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations. *Biostatistics*, **7**: 399–421.

Fedorowicz G, Guerrero S, Wu TD, and Modrusan Z (2009). Microarray analysis of RNA extracted from formalin-fixed, paraffin-embedded and matched fresh-frozen ovarian adenocarcinomas. *BMC Med Genomics*, **2**: 23.

Frank M, Döring C, Metzler D, Eckerle S, and Hansmann ML (2007). Global gene expression profiling of formalin-fixed paraffin-embedded tumor samples: a comparison to snap-frozen material using oligonucleotide microarrays. *Virchows Arch*, **450**: 699–711.

Fridlyand J, Snijders AM, Ylstra B, Li H, Olshen A, Segraves R, Dairkee S, Tokuyasu T, Ljung BM, Jain AN, McLennan J, *et al.* (2006). Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer*, **6**: 96.

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, and Stratton MR (2004). A census of human cancer genes. *Nat Rev Cancer*, **4**: 177–183.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, *et al.* (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**: R80.

Gilbert N and Allan J (2014). Supercoiling in DNA and chromatin. *Curr Opin Genet Dev*, **25**: 15–21.

Glenn TC (2011). Field guide to next-generation DNA sequencers. *Mol Ecol Resour*, **11**: 759–769.

González JR, Subirana I, Escaramís G, Peraza S, Cáceres A, Estivill X, and Armengol L (2009). Accounting for uncertainty when assessing association between copy number and disease: a latent class model. *BMC Bioinformatics*, **10**: 172.

Guo Y, Sheng Q, Samuels DC, Lehmann B, Bauer JA, Pietenpol J, and Shyr Y (2013). Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. *Biomed Res Int*, **2013**: 915636.

Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, and Berri S (2012). Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics*, **28**: 40–47.

Ha G, Roth A, Lai D, Bashashati A, Ding J, Goya R, Giuliany R, Rosner J, Oloumi A, Shumansky K, Chin SF, *et al.* (2012). Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res*, **22**: 1995–2007.

Habermann JK, Doering J, Hautaniemi S, Roblick UJ, Bündgen NK, Nicorici D, Kronenwett U, Rathnagiriswaran S, Mettu RKR, Ma Y, Krüger S, *et al.* (2009). The gene expression signature of genomic instability in breast cancer is an independent predictor of clinical outcome. *Int J Cancer*, **124**: 1552–1564.

Hajirasouliha I, Hormozdiari F, Alkan C, Kidd JM, Birol I, Eichler EE, and Sahinalp SC (2010). Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics*, **26**: 1277–1283.

Hanahan D and Weinberg RA (2011). Hallmarks of cancer: the next generation. *Cell*, **144**: 646–674.

Handsaker RE, Korn JM, Nemesh J, and McCarroll SA (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet*, **43**: 269–276.

Hastings RJ, Bown N, Tibiletti MG, Debiec-Rychter M, Vanni R, Espinet B, van Roy N, Roberts P, van den Berg-de Ruiter E, Bernheim A, Schoumans J, *et al.* (2016). Guidelines for cytogenetic investigations in tumours. *Eur J Hum Genet*, **24**: 6–13.

Hedegaard J, Thorsen K, Lund MK, Hein AMK, Hamilton-Dutoit SJ, Vang S, Nordentoft I, Birkenkamp-Demtröder K, Kruhøffer M, Hager H, Knudsen B, *et al.* (2014). Next-generation sequencing of RNA and DNA isolated from paired fresh-frozen and formalin-fixed paraffin-embedded samples of human cancer and normal tissue. *PLoS One*, **9**: e98187.

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MDM, Niu B, McLellan MD, Uzunangelov V, Zhang J, *et al.* (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, **158**: 929–944.

Hormozdiari F, Alkan C, Eichler EE, and Sahinalp SC (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res*, **19**: 1270–1278.

Hsu L, Self SG, Grove D, Randolph T, Wang K, Delrow JJ, Loo L, and Porter P (2005). Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, **6**: 211–226.

Hu X, Yuan J, Shi Y, Lu J, Liu B, Li Z, Chen Y, Mu D, Zhang H, Li N, Yue Z, *et al.* (2012). pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics*, **28**: 1533–1535.

Huang W, Li L, Myers JR, and Marth GT (2012). ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**: 593–594.

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, *et al.* (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods*, **12**: 115–121.

Hupé P, Stransky N, Thiery JP, Radvanyi F, and Barillot E (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, **20**: 3413–3422.

van den IJssel P, Tijssen M, Chin SF, Eijk P, Carvalho B, Hopmans E, Holstege H, Bangarusamy DK, Jonkers J, Meijer GA, Caldas C, *et al.* (2005). Human and mouse oligonucleotide-based array CGH. *Nucleic Acids Res*, **33**: e192.

International HapMap Consortium (2003). The International HapMap Project. *Nature*, **426**: 789–796.

Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, *et al.* (2009). Repeatability of published microarray gene expression analyses. *Nat Genet*, **41**: 149–155.

Iqbal Z, Caccamo M, Turner I, Flicek P, and McVean G (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet*, **44**: 226–232.

Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, Shafi S, Johnson DH, Mitter R, Rosenthal R, Salm M, *et al.* (2017). Tracking the evolution of non-small-cell lung cancer. *N Engl J Med*, **376**: 2109–2121.

Jong K, Marchiori E, Meijer G, Vaart AVD, and Ylstra B (2004). Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, **20**: 3636–3637.

Jong K, Marchiori E, van der Vaart A, Chin SF, Carvalho B, Tijssen M, Eijk PP, van den Ijssel P, Grabsch H, Quirke P, Oudejans JJ, *et al.* (2007). Cross-platform array comparative genomic hybridization meta-analysis separates hematopoietic and mesenchymal from epithelial tumors. *Oncogene*, **26**: 1499–1506.

Jöreskog KG and Moustaki I (2001). Factor analysis of ordinal variables: a comparison of three apporaches. *Multivar Behav Res*, **36**: 347–387.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, and Pinkel D (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**: 818–821.

Kent WJ (2002). BLAT–the BLAST-like alignment tool. *Genome Res*, **12**: 656–664.

Khojasteh M, Lam WL, Ward RK, and MacAulay C (2005). A stepwise framework for the normalization of array CGH data. *BMC Bioinformatics*, **6**: 274.

Kim S, Jeong K, and Bafna V (2013a). Wessim: a whole-exome sequencing simulator based on in silico exome capture. *Bioinformatics*, **29**: 1076–1077.

Kim TM, Xi R, Luquette LJ, Park RW, Johnson MD, and Park PJ (2013b). Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Res*, **23**: 217–227.

Koboldt DC, Ding L, Mardis ER, and Wilson RK (2010). Challenges of sequencing human genomes. *Brief Bioinform*, **11**: 484–498.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, and Wilson RK (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*, **22**: 568–576.

Kodama Y, Shumway M, Leinonen R, and International Nucleotide Sequence Database Collaboration (2012). The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res*, **40**: D54–D56.

Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, *et al.* (2015). ArrayExpress update–simplifying data submissions. *Nucleic Acids Res*, **43**: D1113–D1116.

Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, and Gerstein MB (2009). PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol*, **10**: R23.

Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, *et al.* (2007). Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**: 420–426.

Krijgsman O, Carvalho B, Meijer GA, Steenbergen RDM, and Ylstra B (2014). Focal chromosomal copy number aberrations in cancer—needles in a genome haystack. *Biochim Biophys Acta*, **1843**: 2698–2704.

Krijgsman O, Israeli D, van Essen HF, Eijk PP, Berens MLM, Mellink CHM, Nieuwint AW, Weiss MM, Steenbergen RDM, Meijer GA, and Ylstra B, *et al.* (2013). Detection limits of DNA copy number alterations in heterogeneous cell populations. *Cell Oncol (Dordr)*, **36**: 27–36.

Krijgsman O, Israeli D, Haan JC, van Essen HF, Smeets SJ, Eijk PP, Steenbergen RDM, Kok K, Tejpar S, Meijer GA, and Ylstra B, *et al.* (2012). CGH arrays compared for DNA isolated from formalin-fixed, paraffin-embedded material. *Genes Chromosomes Cancer*, **51**: 344–352.

Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, NHLBI Exome Sequencing Project, Quinlan AR, Nickerson DA, and Eichler EE (2012). Copy number variation detection and genotyping from exome sequence data. *Genome Res*, **22**: 1525–1532.

Kuilman T, Velds A, Kemper K, Ranzani M, Bombardelli L, Hoogstraat M, Nevedomskaya E, Xu G, de Ruiter J, Lolkema MP, Ylstra B, *et al.* (2015). CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol*, **16**: 49.

Lai WR, Johnson MD, Kucherlapati R, and Park PJ (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**: 3763–3770.

Langmead B and Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**: 357–359.

Langmead B, Trapnell C, Pop M, and Salzberg SL (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, **10**: R25.

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, *et al.* (2011). The European Nucleotide Archive. *Nucleic Acids Res*, **39**: D28–D31.

Li H and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**: 1754–1760.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, *et al.* (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, **20**: 265–272.

Lingjaerde OC, Baumbusch LO, Liestol K, Glad IK, and Borresen-Dale AL (2005). CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics*, **21**: 821–822.

Liu B, Morrison CD, Johnson CS, Trump DL, Qin M, Conroy JC, Wang J, and Liu S (2013). Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget*, **4**: 1868–1881.

Liu J, Mohammed J, Carter J, Ranka S, Kahveci T, and Baudis M (2006). Distance-based clustering of CGH data. *Bioinformatics*, **22**: 1971–1978.

Liu J, Ranka S, and Kahveci T (2007). Markers improve clustering of CGH data. *Bioinformatics*, **23**: 450–457.

Louhimo R, Lepikhova T, Monni O, and Hautaniemi S (2012). Comparative analysis of algorithms for integration of copy number and expression data. *Nat Methods*, **9**: 351–355.

Macintyre G, Ylstra B, and Brenton JD (2016). Sequencing structural variants in cancer for precision therapeutics. *Trends Genet*, **32**: 530–542.

Magi A, Benelli M, Yoon S, Roviello F, and Torricelli F (2011). Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic Acids Res*, **39**: e65.

Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, Battaglia C, Bonora E, Kurg A, Seri M, Magini P, *et al.* (2013). EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol*, **14**: R120.

Magi A, Tattini L, Pippucci T, Torricelli F, and Benelli M (2012). Read count approach for DNA copy number variants detection. *Bioinformatics*, **28**: 470–478.

Majewski IJ, Mittempergher L, Davidson NM, Bosma A, Willems SM, Horlings HM, de Rink I, Greger L, Hooijer GKJ, Peters D, Nederlof PM, *et al.* (2013). Identification of recurrent FGFR3 fusion genes in lung cancer through kinome-centred RNA sequencing. *J Pathol*, **230**: 270–276.

Mann HB and Whitney DR (1947). On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat*, **18**.

Marioni JC, Thorne NP, and Tavare S (2006). BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22**: 1144–1146.

Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, Andrews TD, Stranger BE, Lynch AG, Dermitzakis ET, Carter NP, *et al.* (2007). Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol*, **8**: R228.

Mayrhofer M, DiLorenzo S, and Isaksson A (2013). Patchwork: allele-specific copy number analysis of whole-genome sequenced tumor tissue. *Genome Biol*, **14**: R24.

McElroy KE, Luciani F, and Thomas T (2012). GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics*, **13**: 74.

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC, Zhang Z, *et al.* (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res*, **19**: 1527–1541.

Medvedev P, Fiume M, Dzamba M, Smith T, and Brudno M (2010). Detecting copy number variation with mated short reads. *Genome Res*, **20**: 1613–1622.

Medvedev P, Stanciu M, and Brudno M (2009). Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods*, **6**: S13–S20.

Meyerson M, Gabriel S, and Getz G (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, **11**: 685–96.

Miller CA, Hampton O, Coarfa C, and Milosavljevic A (2011). ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One*, **6**: e16327.

Myers CL, Dunham MJ, Kung SY, and Troyanskaya OG (2004). Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, **20**: 3533–3543.

Myllykangas S, Himberg J, Bohling T, Nagy B, Hollmen J, and Knuutila S (2006). DNA copy number amplification profiling of human neoplasms. *Oncogene*, **25**: 7324–7332.

Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, *et al.* (2011). Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res*, **39**: e90.

Neuvial P, Hupé P, Brito I, Liva S, Manié E, Brennetot C, Radvanyi F, Aurias A, and Barillot E (2006). Spatial normalization of array-CGH data. *BMC Bioinformatics*, **7**: 264.

Nijkamp JF, van den Broek MA, Geertman JMA, Reinders MJT, Daran JMG, and de Ridder D (2012). De novo detection of copy number variation by co-assembly. *Bioinformatics*, **28**: 3195–3202.

Nord AS, Lee M, King MC, and Walsh T (2011). Accurate and exact CNV identification from targeted high-throughput sequence data. *BMC Genomics*, **12**: 184.

Nymark P, Wikman H, Ruosaari S, Hollmen J, Vanhala E, Karjalainen A, Anttila S, and Knuutila S (2006). Identification of specific gene copy number changes in asbestos-related lung cancer. *Cancer Res*, **66**: 5737–5743.

Olshen AB, Bengtsson H, Neuvial P, Spellman PT, Olshen RA, and Seshan VE (2011). Parent-specific copy number in paired tumor-normal studies using circular binary segmentation. *Bioinformatics*, **27**: 2038–2046.

Olshen AB, Venkatraman ES, Lucito R, and Wigler M (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**: 557–572.

Ortiz-Estevez M, Aramburu A, Bengtsson H, Neuvial P, and Rubio A (2012). CalMaTe: a method and software to improve allele-specific copy number of SNP arrays for downstream segmentation. *Bioinformatics*, **28**: 1793–1794.

Picard F, Robin S, Lavielle M, Vaisse C, and Daudin JJ (2005). A statistical approach for array CGH data analysis. *BMC Bioinformatics*, **6**: 27.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, *et al.* (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, **20**: 207–211.

Pinto D, Darvishi K, Shi X, Rajan D, Rigler D, Fitzgerald T, Lionel AC, Thiruvahindrapuram B, Macdonald JR, Mills R, Prasad A, *et al.* (2011). Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol*, **29**: 512–520.

Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, Williams CF, Jeffrey SS, Botstein D, and Brown PO (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet*, **23**: 41–46.

Pop M, Phillippy A, Delcher AL, and Salzberg SL (2004). Comparative genome assembly. *Brief Bioinform*, **5**: 237–248.

Popova T, Manié E, Stoppa-Lyonnet D, Rigaill G, Barillot E, and Stern MH (2009). Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. *Genome Biol*, **10**: R128.

Qi J and Zhao F (2011). inGAP-sv: a novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Res*, **39**: W567–W575.

Quinlan AR, Clark RA, Sokolova S, Leibowitz ML, Zhang Y, Hurles ME, Mell JC, and Hall IM (2010). Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Res*, **20**: 623–635.

Rancoita PMV, Hutter M, Bertoni F, and Kwee I (2009). Bayesian DNA copy number analysis. *BMC Bioinformatics*, **10**: 10.

Rancoita PMV, Hutter M, Bertoni F, and Kwee I (2010). An integrated bayesian analysis of LOH and copy number data. *BMC Bioinformatics*, **11**: 321.

Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, and Smyth GK (2007). A comparison of background correction methods for two-colour microarrays. *Bioinformatics*, **23**: 2700–2707.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, and Jaffe DB (2013). Characterizing and measuring bias in sequence data. *Genome Biol*, **14**: R51.

Rouveirol C, Stransky N, Hupé P, Rosa PL, Viara E, Barillot E, and Radvanyi F (2006). Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, **22**: 849–856.

Rueda OM and Diaz-Uriarte R (2009). Detection of recurrent copy number alterations in the genome: taking among-subject heterogeneity seriously. *BMC Bioinformatics*, **10**: 308.

Russnes HG, Vollan HKM, Lingjaerde OC, Krasnitz A, Lundin P, Naume B, Sørlie T, Borgen E, Rye IH, Langerød A, Chin SF, *et al.* (2010). Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci Transl Med*, **2**: 38ra47.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, and Erlich HA (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**: 487–491.

Santarius T, Shipley J, Brewer D, Stratton MR, and Cooper CS (2010). A census of amplified and overexpressed human cancer genes. *Nat Rev Cancer*, **10**: 59–64.

Sathirapongsasuti JF, Lee H, Horst BAJ, Brunner G, Cochran AJ, Binder S, Quackenbush J, and Nelson SF (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics*, **27**: 2648–2654.

Scheinin I, Ferreira JA, Knuutila S, Meijer GA, van de Wiel MA, and Ylstra B (2010). CGHpower: exploring sample size calculations for chromosomal copy number experiments. *BMC Bioinformatics*, **11**: 331–340.

Scheinin I, Myllykangas S, Borze I, Böhling T, Knuutila S, and Saharinen J (2008). CanGEM: mining gene copy number changes in cancer. *Nucleic Acids Res*, **36**: D830–D835.

Scheinin I, Sie D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF, van Essen HF, Eijk PP, Rustenburg F, Meijer GA, Reijneveld JC, *et al.* (2014). DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res*, **24**: 2022–2032.

Schweiger MR, Kerick M, Timmermann B, Albrecht MW, Borodina T, Parkhomchuk D, Zatloukal K, and Lehrach H (2009). Genome-wide massively parallel sequencing of formaldehyde fixed-paraffin embedded (FFPE) tumor tissues for copy-number- and mutation-analysis. *PLoS One*, **4**: e5548.

Scicchitano MS, Dalmas DA, Bertiaux MA, Anderson SM, Turner LR, Thomas RA, Mirable R, and Boyce RW (2006). Preliminary comparison of quantity, quality, and microarray performance of RNA extracted from formalin-fixed, paraffin-embedded, and unfixed frozen tissue samples. *J Histochem Cytochem*, **54**: 1229–1237.

Shah SP (2008). Computational methods for identification of recurrent copy number alteration patterns by array CGH. *Cytogenet Genome Res*, **123**: 343–351.

Shah SP, Cheung KJ Jr, Johnson NA, Alain G, Gascoyne RD, Horsman DE, Ng RT, and Murphy KP (2009). Model-based clustering of array CGH data. *Bioinformatics*, **25**: i30–i38.

Shen JJ and Zhang NR (2012). Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *Ann Appl Stat*, **6**: 476–496.

Shlien A and Malkin D (2010). Copy number variations and cancer susceptibility. *Curr Opin Oncol*, **22**: 55–63.

Sie D, Snijders PJF, Meijer GA, Doeleman MW, van Moorsel MIH, van Essen HF, Eijk PP, Grünberg K, van Grieken NCT, Thunnissen E, Verheul HM, *et al.* (2014). Performance of amplicon-based next generation DNA sequencing for diagnostic gene mutation profiling in oncopathology. *Cell Oncol (Dordr)*, **37**: 353–361.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, and Birol I (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res*, **19**: 1117–1123.

Sindi SS, Onal S, Peng LC, Wu HT, and Raphael BJ (2012). An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol*, **13**: R22.

Smeets SJ, Harjes U, van Wieringen WN, Sie D, Brakenhoff RH, Meijer GA, and Ylstra B (2011). To DNA or not to DNA? That is the question, when it comes to molecular subtyping for the clinic! *Clin Cancer Res*, **17**: 4959–4964.

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, *et al.* (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet*, **29**: 263–264.

Sokolova V, Crippa E, and Gariboldi M (2016). Integration of genome scale data for identifying new players in colorectal cancer. *World J Gastroenterol*, **22**: 534–545.

Somiari SB, Shriver CD, He J, Parikh K, Jordan R, Hooke J, Hu H, Deyarmin B, Lubert S, Malicki L, Heckman C, *et al.* (2004). Global search for chromosomal abnormalities in infiltrating ductal carcinoma of the breast using array-comparative genomic hybridization. *Cancer Genet Cytogenet*, **155**: 108–118.

Staaf J, Jönsson G, Ringnér M, and Vallon-Christersson J (2007). Normalization of array-CGH data: influence of copy number imbalances. *BMC Genomics*, **8**: 382.

Staden R (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res*, **6**: 2601–2610.

Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, *et al.* (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**: 75–81.

Talevich E, Shain AH, Botton T, and Bastian BC (2016). CNVkit: Genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput Biol*, **12**: e1004873.

Taub MA, Corrada Bravo H, and Irizarry RA (2010). Overcoming bias and systematic errors in next generation sequencing data. *Genome Med*, **2**: 87.

Teo SM, Pawitan Y, Ku CS, Chia KS, and Salim A (2012). Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, **28**: 2711–2718.

van Thuijl HF, Scheinin I, Sie D, Alentorn A, van Essen HF, Cordes M, Fleischeuer R, Gijtenbeek AM, Beute G, van den Brink WA, Meijer GA, *et al.* (2014). Spatial and temporal evolution of distal 10q deletion, a prognostically unfavorable event in diffuse low-grade gliomas. *Genome Biol*, **15**: 471–483.

Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, and Altman RB (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**: 520–525.

Tsafrir D, Bacolod M, Selvanayagam Z, Tsafrir I, Shia J, Zeng Z, Liu H, Krier C, Stengel RF, Barany F, Gerald WL, *et al.* (2006). Relationship of gene expression and chromosomal abnormalities in colorectal cancer. *Cancer Res*, **66**: 2129–2137.

Unger K, Malisch E, Thomas G, Braselmann H, Walch A, Jackl G, Lewis P, Lengfelder E, Bogdanova T, Wienberg J, and Zitzelsberger H, *et al.* (2008). Array CGH demonstrates characteristic aberration signatures in human papillary thyroid carcinomas governed by RET/PTC. *Oncogene*, **27**: 4592–4602.

Varambally S, Cao Q, Mani RS, Shankar S, Wang X, Ateeq B, Laxman B, Cao X, Jing X, Ramnarayanan K, Brenner JC, *et al.* (2008). Genomic loss of microRNA-101 leads to overexpression of histone methyltransferase EZH2 in cancer. *Science*, **322**: 1695–9.

Venkatraman ES and Olshen AB (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**: 657–663.

Wang P (2009). Algorithms for calling gains and losses in array CGH data. *Methods Mol Biol*, **556**: 99–116.

Wang P, Kim Y, Pollack J, Narasimhan B, and Tibshirani R (2005). A method for calling gains and losses in array CGH data. *Biostatistics*, **6**: 45–58.

Wang XV, Blades N, Ding J, Sultana R, and Parmigiani G (2012). Estimation of sequencing error rates in short reads. *BMC Bioinformatics*, **13**: 185.

Whiteford N, Haslam N, Weber G, Prügel-Bennett A, Essex JW, Roach PL, Bradley M, and Neylon C (2005). An analysis of the feasibility of short read sequencing. *Nucleic Acids Res*, **33**: e171.

van de Wiel MA, Brosens R, Eilers PHC, Kumps C, Meijer GA, Menten B, Sistermans E, Speleman F, Timmerman ME, and Ylstra B (2009). Smoothing waves in array CGH tumor profiles. *Bioinformatics*, **25**: 1099–1104.

van de Wiel MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, and Ylstra B (2007). CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, **23**: 892–894.

van de Wiel MA, Picard F, van Wieringen WN, and Ylstra B (2011). Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief Bioinform*, **12**: 10–21.

van de Wiel MA, Smeets SJ, Brakenhoff RH, and Ylstra B (2005). CGHMultiArray: exact P-values for multi-array comparative genomic hybridization data. *Bioinformatics*, **21**: 3193–3194.

van de Wiel MA and van Wieringen WN (2007). CGHregions: dimension reduction for array CGH data with minimal information loss. *Cancer Informatics*, **3**: 55–63.

van Wieringen WN, van de Wiel MA, and van der Vaart AW (2008a). A test for partial differential expression. *J Amer Statist Assoc*, **103**: 1039–1049.

van Wieringen WN, van de Wiel MA, and Ylstra B (2007). Normalized, segmented or called aCGH data? *Cancer Inform*, **3**: 321–327.

van Wieringen WN, van de Wiel MA, and Ylstra B (2008b). Weighted clustering of called array CGH data. *Biostatistics*, **9**: 484–500.

Wilhelm M, Veltman JA, Olshen AB, Jain AN, Moore DH, Presti JC Jr, Kovacs G, and Waldman FM (2002). Array-based comparative genomic hybridization for the differential diagnosis of renal cell cancer. *Cancer Res*, **62**: 957–960.

Willenbrock H and Fridlyand J (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**: 4084–4091.

Wood HM, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, McKinley C, Egan P, Ross L, Hayward B, Morgan J, *et al.* (2010). Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of DNA from formalin-fixed paraffin-embedded specimens. *Nucleic Acids Res*, **38**: e151.

Xi R, Hadjipanayis AG, Luquette LJ, Kim TM, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA, Kucherlapati R, *et al.* (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci U S A*, **108**: E1128–E1136.

Xie C and Tammi MT (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, **10**: 80.

Yau C (2013). OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics*, **29**: 2482–2484.

Ye K, Schulz MH, Long Q, Apweiler R, and Ning Z (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**: 2865–2871.

Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, and Meijer GA (2006). BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res*, **34**: 445–450.

Yoon S, Xuan Z, Makarov V, Ye K, and Sebat J (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res*, **19**: 1586–1592.

Zeitouni B, Boeva V, Janoueix-Lerosey I, Loeillet S, Legoix-né P, Nicolas A, Delattre O, and Barillot E (2010). SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, **26**: 1895–6.

Zerbino DR and Birney E (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, **18**: 821–9.

Zhang H, Liu T, Zhang Z, Payne SH, Zhang B, McDermott JE, Zhou JY, Petyuk VA, Chen L, Ray D, Sun S, *et al.* (2016). Integrated proteogenomic characterization of human high-grade serous ovarian cancer. *Cell*, **166**: 755–765.

Zhang J and Wu Y (2011). SVseq: an approach for detecting exact breakpoints of deletions with low-coverage sequence data. *Bioinformatics*, **27**: 3228–3234.

Zhang NR, Senbabaoglu Y, and Li JZ (2010). Joint estimation of DNA copy number from multiple platforms. *Bioinformatics*, **26**: 153–160.

Zhao M, Wang Q, Wang Q, Jia P, and Zhao Z (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, **14 Suppl 11**: S1.

Zhu M, Need AC, Han Y, Ge D, Maia JM, Zhu Q, Heinzen EL, Cirulli ET, Pelak K, He M, Ruzzo EK, *et al.* (2012). Using ERDS to infer copy-number variants in high-coverage genomes. *Am J Hum Genet*, **91**: 408–421.

Chapter 2

CanGEM: mining gene copy number changes in cancer

Ilari Scheinin
Samuel Myllykangas
Ioana Borze
Tom Böhling
Sakari Knuutila
Juha Saharinen

Nucleic Acids Research (2008) **36:** D830–D835

D830–D835 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 11 October 2007*
doi:10.1093/nar/gkm802

CanGEM: mining gene copy number changes in cancer

Ilari Scheinin[1,2,3], Samuel Myllykangas[3], Ioana Borze[3], Tom Böhling[3], Sakari Knuutila[3] and Juha Saharinen[1,2,*]

[1]Genome Informatics Unit, Biomedicum Helsinki, Finland, [2]Department of Molecular Medicine, National Public Health Institute of Finland, KTL and [3]Departments of Pathology, Haartman Institute and HUSLAB, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland

Received August 15, 2007; Revised September 16, 2007; Accepted September 17, 2007

ABSTRACT

The use of genome-wide and high-throughput screening methods on large sample sizes is a well-grounded approach when studying a process as complex and heterogeneous as tumorigenesis. Gene copy number changes are one of the main mechanisms causing cancerous alterations in gene expression and can be detected using array comparative genomic hybridization (aCGH). Microarrays are well suited for the integrative systems biology approach, but none of the existing microarray databases is focusing on copy number changes. We present here CanGEM (Cancer GEnome Mine), which is a public, web-based database for storing quantitative microarray data and relevant metadata about the measurements and samples. CanGEM supports the MIAME standard and in addition, stores clinical information using standardized controlled vocabularies whenever possible. Microarray probes are re-annotated with their physical coordinates in the human genome and aCGH data is analyzed to yield gene-specific copy numbers. Users can build custom datasets by querying for specific clinical sample characteristics or copy number changes of individual genes. Aberration frequencies can be calculated for these datasets, and the data can be visualized on the human genome map with gene annotations. Furthermore, the original data files are available for more detailed analysis. The CanGEM database can be accessed at http://www.cangem.org/.

INTRODUCTION

With the exception of few hematologic malignancies that are characterized by a single chromosomal change, e.g. the BCR-ABL1 fusion gene in chronic myelogenous leukemia (1), cancer is generally a complex disease. Especially carcinomas, which usually undergo prolonged carcinogenesis and account for over 80% of cancer-related deaths (2), are usually characterized by chaotic genomic changes. Chromosomal or gene copy number alterations are one of the most important mechanisms that perturb normal gene function by inducing changes reflected in gene expression (3). Other types of changes include mutations (inherited or somatic), translocations and changes in epigenetic make-up that affect the gene regulation machinery and protein structure and function. During carcinogenesis and cancer progression, activation and malfunction of a number of cancer genes are required for cancer cells to gain the independence of growth supporting signaling and immunity to growth restrains, evade apoptosis and replicate unlimitedly, sustain angiogenesis and escape the control of the anatomical primary site, thus, acquire the hallmarks of cancer (4). In addition to cancer, gene copy number aberrations are also important in many congenital disorders, especially small deletions that are detectable using high-resolution aCGH.

Array comparative genomic hybridization (aCGH) is a technique that uses microarrays to detect changes in gene copy number, and is widely used in cancer research to characterize different tumors and hematologic malignancies (5). Arrays can be manufactured using different techniques, and recently synthetic oligonucleotides have been gaining popularity from spotted BAC or cDNA clones (6). As the selection of used arrays is wide, it is necessary to be able to integrate data measured with different platforms, in order to be able to do large-scale studies. This is further emphasized by the in-house manufactured spotted arrays.

Copy number aberrations comprise of deletions and amplifications, which promote cancer by acting on tumor suppressor genes and proto-oncogenes, respectively. These genes can be commonly termed as 'cancer genes'. Because aberrations are formed through a process of common

*To whom correspondence should be addressed. Tel: +358 9 4744 8969; Fax: +358 9 4744 8480; Email: juha.saharinen@ktl.fi

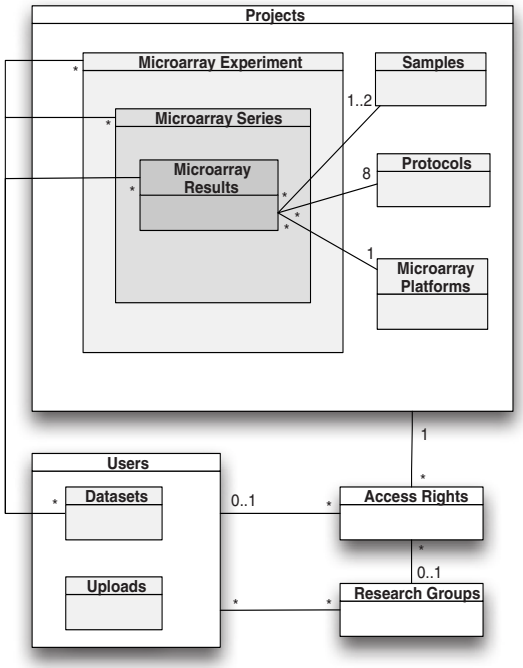

Figure 1. Database structure. This figure summarizes the relationships between the different data entities that are used in the database. Microarray results are obtained from a single microarray hybridization and contain a text file with a numerical representation of the measured spot intensities obtained from the scanned array with an image analysis software. It can also include the image file itself. In addition to these files, results contain links to the biological specimens (samples), experimental procedures (protocols) and the specific microarray platform that were used to obtain the results. The protocols section is divided into eight different stages: extraction, digestion, amplification, labeling, hybridization, washing, scanning and image analysis. Together they correspond to the methods section of an article preceding the data analysis stage. Sample and protocol information is submitted to the database separately from the microarray results to allow the reuse of the same samples and protocols for multiple hybridizations. An example is a study that integrates the results of multiple array techniques, such as both copy number and expression data. A number of results can be combined into a series, and multiple series can be further combined to form an experiment, which corresponds to a published article. All of the data entities mentioned above are contained within projects, which allow user permissions to be specified on a per user account or per research group basis. The service can therefore be used to aid data sharing between collaborators in preliminary prepublication stages, or to give access to manuscript referees. Even though this could also allow the users to continue to limit the availability of their data, everything uploaded to the CanGEM database should be made publicly available once the researchers' get their results published. There are also two data types that are user-account specific: uploads and datasets. They are only visible to that specific user account. Uploads are files (e.g. microarray result files) that have been uploaded to the web server, but not yet used to create an actual database entry. Datasets are user-defined collections of microarray data, and can be constructed manually or as saved search queries. These smart datasets get updated automatically and can be configured to send email alerts when their contents change, i.e. when new microarray data become available that match previously defined search criteria, e.g. of tissue type, cancer type and age group of interest. The difference between datasets and microarray results, series and experiments, is that the latter ones are defined by the original submitter and are the same for everybody, while every user can create custom datasets to meet their specific needs. *, Asterisk represent the numbers next to the lines connecting the boxes describe the relationship between the two data entities. For example, each microarray result is linked to either one or two samples depending on the array type, and this is denoted with 1..2. Each sample can be used for an arbitrary number of microarray results, which is depicted with the symbol.

breakpoint errors in DNA replication and/or repair followed by natural selection, the altered genomic regions usually contain not only cancer genes, but also bystanders. Identification of the driving genes is essential in understanding cancer biology as well as for clinical applications, namely, prognostics, diagnostics and therapeutics, where specific targets are sought after. Because of the complexity of carcinogenesis and the nature of the process creating aberrations, large-scale screening studies are needed to achieve this goal.

The existing microarray databases [such as ArrayExpress (7), Gene Expression Omnibus (8), and Stanford MicroArray Database (9)] are focused on gene expression data and do not provide tools for studying copy number changes. We present here CanGEM (Cancer GEnome Mine), which is a public, web-based database service for storing clinical information about tumor samples and related microarray data. Emphasis is on copy number changes, but also other types of microarray data can be stored, including locus heterogeneity and gene expression data, typically when collected from the same samples.

DESIGN, IMPLEMENTATION AND USAGE OF CANGEM

Database structure

The structure of the CanGEM database is MIAME-compliant (10) and flexible in allowing the storage of different file formats from different software packages. Figure 1 summarizes the relationships between different data entities that are used in this article to describe the database.

Annotation of samples

In order to support systematic research, samples in CanGEM are annotated using classification systems based on controlled vocabularies, instead of free text descriptions common in many gene expression microarray databases. To describe the topographical and morphological attributes of the cancer, we are using chapter II (Neoplasms) of the International Statistical Classification of Diseases and Related Health Problems (10th Revision; ICD-10) and International Classification of Disease for Oncology (3rd Edition; ICD-O-3) of the World Health Organization (WHO). They are both three-step hierarchical ontologies that allow a precise classification of the cancer types and morphologies. If it is not possible to classify a particular sample up to the most detailed level, the definition can be left at the previous, broader stage. It is also possible to assign multiple definitions to a single sample.

We have also adopted the classification system of the eVOC Ontology, which is a set of ontologies developed

to classify gene expression libraries (11). It contains a total of 10 categories, of which we are using six: anatomical system, cell type, development stage, pathology, tissue preparation and treatment. eVOC provides hierarchical classification systems with a varying number of levels. As with ICD, it is possible to assign multiple ontology terms to a single sample.

For classifying the stage of the cancer, we are using the TNM Classification, which is a systematic way of describing the size and spread of a tumor. It uses three values to describe the size of the primary tumor (T), and whether the cancer has spread to the lymph nodes (N), or to more distant locations in the body (M).

In addition to these classification systems, we are also collecting information about exposure to environmental risk factors, patient sex, tumor size, survival, cause of death, and whether surgery has been done and if it was curative or not. All of the clinical attributes as well as the ICD and eVOC ontologies can be used to search for samples in the database.

Our current sample annotations are rather cancer-specific, but in the future we will be updating the system to better account for other possible uses of array CGH.

Annotation of microarray platforms

Different microarray platforms contain different sets of probes identified by different sets of IDs, which makes it difficult to integrate data from multiple sources. When working with array CGH, this problem is further complicated by the fact that the technique is not gene-centric. BAC arrays, still used for CGH, contain probes that can be 300 kb long and contain several genes. Oligo arrays on the other hand can contain probes that have been specifically designed to match regions between genes. To overcome this problem, we are using probe sequences to re-annotate the microarray probes to physical coordinates of the human genome with a custom iterative algorithm based on MegaBlast (12). Further, this approach enables CanGEM to unite aCGH, LOH and gene expression data through physical coordinates in the future.

In the case of oligonucleotide arrays, the entire sequence of the probe is known, and it often corresponds to an unambiguous and continuous chromosomal sequence. However, with cDNA probes, in most cases, intronic regions split probes to more than one matching exons resulting in multiple blast hits. Also, in the case of cDNA or BAC arrays, the probes are generally too long to be sequenced entirely, and the sequenced sections are from the ends of the probe (preferably from both). After all the sequences for a specific probe have been analyzed, CanGEM joins the mapping results together if all hits are from the same chromosome, and the entire length of the joined probe does not exceed 2.5 Mbp of genomic DNA (the longest human gene in Ensembl 45 is 2 304 117 base pairs). If these conditions are not met, the probe is marked as ambiguous and excluded from further analysis. The probe-to-genome mapping results are saved to CanGEM and used for analysis of submitted microarray data, and the mapping procedure is repeated when a new build of the human genome becomes available. Currently, all the microarray platforms available in CanGEM have been mapped to both NCBI builds 35 and 36. The mapping process is illustrated in Figure 2A.

Currently, the two-color platforms suitable for array CGH in CanGEM include cDNA and oligo-based arrays from Agilent Technologies (12K cDNA, 22K oligos, 44K oligos for CGH and expression, 244K oligo CGH) and one custom cDNA array. Supported one-color expression platforms include Affymetrix U133 and U95 arrays. Furthermore, introduction of new array platforms is done easily.

Submission of data to CanGEM

CanGEM is an open and publicly accessible data repository. However, in order to submit data to CanGEM, users need to register for a (free) user account, which then becomes the owner of that data. Currently, all data submission operations are done through a web browser, and the different data entities are created by filling out simple web forms. For submissions of larger sets of data, a batch tool is available for inserting multiple microarray results. Also, new samples can be created using an existing one as a template speeding up the process, as samples in a single microarray experiment usually share similar characteristics. Step-by-step documentation of the submission process can be found from the CanGEM website. The availability of the submitted data can be controlled as explained in more detail in the legend of Figure 1.

This project has been reviewed and approved by the Ethical Review Board of Helsinki and Uusimaa Hospital District and authorized by the Clinical Review Board of Helsinki University Central Hospital. For ethical reasons, it is forbidden to store any information that could identify the individual patients in CanGEM. This includes any kind of identification codes, but can also encompass extremely rare clinical cases, whose uniqueness could be individualizing. It is the responsibility of the user submitting data to ensure that this requirement is fulfilled, and therefore users willing to submit data to CanGEM have to get a free user id and password.

Processing of the data submitted to CanGEM

After a microarray experiment is uploaded to the database, the data are analyzed using a predefined procedure. The process is semi-automated, involving the data being checked by a human curator. This only includes technical details, such as ensuring that the sample attributes are in place, but does not cover quality control of the arrays, which is left to the user. The details of the analysis algorithm for the R statistical analysis environment can be found from the library package provided on the CanGEM website. In brief, the data are filtered for outliers, background-corrected and lowess normalized in R/Bioconductor using the limma package (13). As the normalization procedure is the same for all different microarray platforms, we have chosen not to use methods that depend on the array design, such as print-tip lowess.

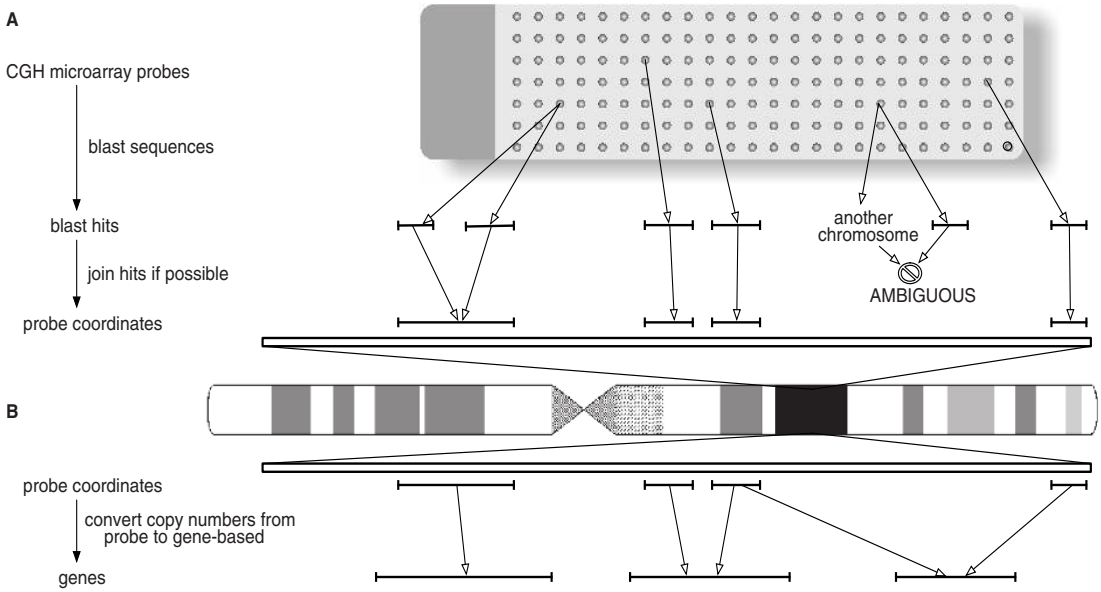

Figure 2. (**A**) Mapping probes to physical coordinates of the genome. First, all available sequences for a specific probe are analyzed with MegaBlast, and the results are joined together if they meet the conditions outlined in the main text. The figure shows this process for five probes on a CGH microarray. Probe 1 yields two blast hits, which are joined together to get the coordinates for that probe. Probes 2, 3 and 5 only produce single hits. Probe 4 gives two matches that are in different chromosomes, and the probe is therefore marked as ambiguous and excluded. (**B**) Converting probe-based data to gene copy numbers. The physical coordinates of the microarray probes, obtained through the predone probe-to-genome mapping process for the used array platform, are used to convert probe-based copy number data to gene-centric. The image shows three genes in this genomic region. The position of gene 1 overlaps with probe 1 on the array, so the copy number of gene 1 is the same as the copy number of probe 1. Gene 2 has two overlapping probes (2 and 3), so its copy number is calculated from these two probes. Gene 3 has no overlapping probes, so its value is derived from the last preceding probe (3) and the first one tailing the gene (probe 5). If the copy number for a gene is calculated from multiple probes, and all these probes share the same value (-1, 0 or $+1$), the gene will receive the same value. If the probes have different values, the gene will be assigned a normal, or unchanged, copy number (0).

Different normalization schemes for array CGH data have been compared in (14) and of the included methods that are not dependent on array design, lowess was found to perform best in removing technical bias, while maintaining the biological significance. After normalization, the data is combined with the physical coordinates for the specific microarray platform precalculated with the probe-to-genome analysis, and all the probes that have been marked as ambiguous or did not produce any hits are removed. Also, data from chromosome Y is removed if the patient is not a male and chromosomes X and Y if the sex does not match with the reference sample. The CGH profiles are calculated using the ACE algorithm of the CGH Explorer program (15), which converts the log ratios to discrete levels of normal, amplified, or deleted copy number, represented with the numbers 0, 1 and -1, respectively. The algorithm also calculates estimates of false discovery rates, and the 'medium' option is selected from the presented alternatives balancing between sensitivity and specificity.

Because different array platforms contain different probes targeting different areas of the genome, the results are converted to gene-specific copy numbers as follows. For each gene, it is first checked if there are probes on the array that overlap with the position of the gene, in which case the copy number for that gene is calculated from the overlapping probes. If there are no such probes, the copy number is calculated from the values of the last preceding probe and the first one tailing the gene. If the probes in question all have the same value (-1; 0; $+1$), then that value is chosen for the gene. If they have different values, the gene gets a copy number of 0, meaning normal or unchanged, state. This process is illustrated in Figure 2B. The gene-specific copy numbers are then stored in the database to allow searching, and calculations of aberration frequencies.

Browsing and searching CanGEM data

Data in CanGEM can be searched either using a free-text search box on the front page of the service, or by using a detailed search form for complex queries using the clinical attributes used to describe samples. This also includes the hierarchical classification systems of ICD and eVOC, in which case a search for a more generic term (e.g. digestive organs) also returns all the results that have been mapped to a child of that ontology term (e.g. stomach and colon). It is also possible to search using the copy number status of a given gene. Further, search results can be saved as datasets for future reference and gene aberration frequencies can be calculated and visualized for these subsets of CanGEM data. Datasets that have been created by saving search results

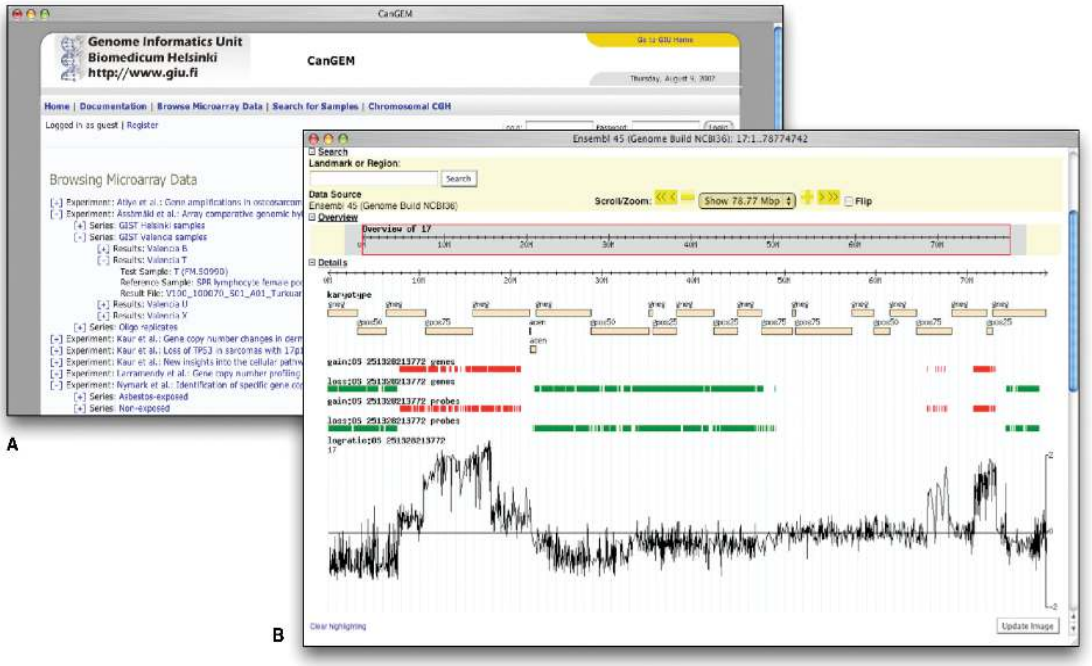

Figure 3. (**A**) Browsing interface. A hierarchical user interface is provided for accessing microarray data. (**B**) Data visualization. The GBrowse software package showing both gene and probe-based copy number aberrations and also the original probe log ratios.

will be automatically updated when new data matching the search conditions becomes available, and email alerts can be sent for such changes. Different microarray data entities are shown using a hierarchical browsing interface as shown in Figure 3.

Data visualization

CanGEM provides an integrated data visualization engine, based on GBrowse (16) and Dazzle/DAS (17) packages. The selected samples are shown with the CGH copy number data on the human genome, together with other annotations from a local installation of the Ensembl database. The user can zoom into a particular region of interest and select the preferred sets of annotations to be displayed. The visualization system provides a quick and easy way to see chromosomal aberrations in the regions of interest and an example of the output is shown in Figure 3. For individual samples, the user can choose to display log ratios and/or probe or gene-based copy numbers. For a collection of samples, the plot shows the frequencies of gains and losses.

It is also possible for users to visualize their own private annotations together with data from CanGEM, or to use another visualization agent that supports the GFF file format.

Retrieval of data from CanGEM

All the data submitted and made publicly available can be downloaded, together with the original raw data files as well as the CanGEM processed numerical data. This allows e.g. performing custom data analysis using the software package and algorithm of choice. Downloading can be done from the web user interface, or using the provided library package for the R statistical analysis environment. Documentation for the available functions is provided within the package.

DISCUSSION

We have presented here a database service for storing clinical information about tumor samples and microarray data, with emphasis on array CGH. The probes on different microarray platforms are mapped to physical coordinates of the human genome and microarray data are analyzed to yield gene-specific copy numbers facilitating the integration of data measured with different array platforms. Data mining of gene copy number changes provides valuable insight into the extremely complicated process of tumorigenesis, and public databases are a prerequisite for this kind of large-scale analysis. Such an approach is indispensable when trying to find aberrations that correlate with a specific diagnostic, prognostic or therapeutic trait, such as poor prognosis or drug resistance. These features might go unnoticed in individual studies, because of the heterogeneity in the processes of tumor progression and aberration formation, but public databases help to improve the statistical power of such analyses.

ACKNOWLEDGEMENTS

Juri Ahokas, Teemu Perheentupa and Tomi Simonen are acknowledged for their technical expertise in computer and database infrastructure. Genome Informatics Unit, Helsinki is acknowledged for their high performance computing facilities. The article was funded by Sigrid Jusélius Foundation; Academy of Finland (207469); Helsinki University Central Hospital Research Funds (EVO, TYH6229); Finnish Cancer Organizations; Finnish Funding Agency for Technology and Innovation (TEKES, 387/31/05); Foundation for Pediatric Research. Funding to pay the Open Access publication charges for the article was provided by National Public Health Institute of Finland.

Conflict of interest statement. None declared.

REFERENCES

1. Heisterkamp,N., Stam,K., Groffen,J., de Klein,A. and Grosveld,G. (1985) Structural organization of the bcr gene and its role in the Ph' translocation. *Nature*, **315**, 758–761.
2. Ferlay,J., Bray,F., Pisani,P. and Parkin,D. (2004) GLOBOCAN 2002: *Cancer Incidence, Mortality and Prevalence Worldwide. IARC CancerBase No. 5. version 2.0*. IARC Press, Lyon.
3. Albertson,D.G. (2006) Gene amplification in cancer. *Trends Genet.*, **22**, 447–455.
4. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, **100**, 57–70.
5. Pinkel,D. and Albertson,D.G. (2005) Comparative genomic hybridization. *Annu. Rev. Genomics Hum. Genet.*, **6**, 331–354.
6. Ylstra,B., van den Ijssel,P., Carvalho,B., Brakenhoff,R.H. and Meijer,G.A. (2006) BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res.*, **34**, 445–450.
7. Parkinson,H., Sarkans,U., Shojatalab,M., Abeygunawardena,N., Contrino,S., Coulson,R., Farne,A., Lara,G.G., Holloway,E. *et al.* (2005) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **33**, D553–D555.
8. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M. *et al.* (2007) NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res.*, **35**, D760–D765.
9. Demeter,J., Beauheim,C., Gollub,J., Hernandez-Boussard,T., Jin,H., Maier,D., Matese,J.C., Nitzberg,M., Wymore,F. *et al.* (2007) The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res.*, **35**, D766–D770.
10. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
11. Kelso,J., Visagie,J., Theiler,G., Christoffels,A., Bardien,S., Smedley,D., Otgaar,D., Greyling,G., Jongeneel,C.V. *et al.* (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.*, **13**, 1222–1230.
12. Zhang,Z., Schwartz,S., Wagner,L. and Miller,W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.
13. Smyth,G.K. and Speed,T. (2003) Normalization of cDNA microarray data. *Methods*, **31**, 265–273.
14. Khojasteh,M., Lam,W.L., Ward,R.K. and MacAulay,C. (2005) A stepwise framework for the normalization of array CGH data. *BMC Bioinformatics*, **6**, 274.
15. Lingjaerde,O.C., Baumbusch,L.O., Liestol,K., Glad,I.K. and Borresen-Dale,A.-L. (2005) CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics*, **21**, 821–822.
16. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
17. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The distributed annotation system. *BMC Bioinformatics*, **2**, 7.

Chapter 3

CGHpower: exploring sample size calculations for chromosomal copy number experiments

Ilari Scheinin
José A Ferreira
Sakari Knuutila
Gerrit A Meijer
Mark A van de Wiel
Bauke Ylstra

BMC Bioinformatics (2010) **11:** 331–340

BMC Bioinformatics

SOFTWARE Open Access

CGHpower: exploring sample size calculations for chromosomal copy number experiments

Ilari Scheinin[1,2,3], José A Ferreira[4], Sakari Knuutila[2], Gerrit A Meijer[1], Mark A van de Wiel[4,5] and Bauke Ylstra*[1]

Abstract

Background: Determining a suitable sample size is an important step in the planning of microarray experiments. Increasing the number of arrays gives more statistical power, but adds to the total cost of the experiment. Several approaches for sample size determination have been developed for expression array studies, but so far none has been proposed for array comparative genomic hybridization (aCGH).

Results: Here we explore power calculations for aCGH experiments comparing two groups. In a pilot experiment CGHpower estimates the biological diversity between groups and provides a statistical framework for estimating average power as a function of sample size. As the method requires pilot data, it can be used either in the planning stage of larger studies or in estimating the power achieved in past experiments.

Conclusions: The proposed method relies on certain assumptions. According to our evaluation with public and simulated data sets, they do not always hold true. Violation of the assumptions typically leads to unreliable sample size estimates. Despite its limitations, this method is, at least to our knowledge, the only one currently available for performing sample size calculations in the context of aCGH. Moreover, the implementation of the method provides diagnostic plots that allow critical assessment of the assumptions on which it is based and hence on the feasibility and reliability of the sample size calculations in each case.

The CGHpower web application and the program outputs from evaluation data sets can be freely accessed at http://www.cangem.org/cghpower/

Background

Array comparative genomic hybridization (aCGH) is a technique that uses microarrays to perform high-resolution and genome-wide screening of DNA copy number changes. Its most important applications are in cancer research [1] and clinical genetics [2]. In this paper we focus on aCGH experiments comparing two groups of cancer samples. Previously, we introduced the Wilcoxon test with ties to identify chromosomal copy number differences when comparing two groups [3]. The goal of comparing two groups is generally to identify disease biomarkers, chromosomal regions (or genes therein) for survival, therapy, progression, *et cetera*. An important problem that arises in the planning of aCGH experiments is the choice of the sample size, which we explore here. Data analysis of microarray experiments comparing two groups generally involves calculating a test statistic for each array element and setting a cutoff for rejecting the null hypothesis of no difference between the groups. With a single array element, there are therefore two typical errors that can occur in the process. A type I error occurs when the null hypothesis is rejected even though it was actually true and the cut-off was exceeded only by chance. A type II error involves accepting a null hypothesis that should have been rejected, thus failing to identify a true difference. To broaden the perspective from individual array elements to the framework of multiple testing covering the entire microarray, two concepts are used: false discovery rate (FDR) [4] and average power. FDR is the expected percentage of discoveries that are false. Statistical power is the probability of recognizing a single array element with a true difference, and average power refers to the expected percentage of true positives that is identified. In general, it is desirable to have the FDR as close to zero and average power as close to one as possible. Setting the cut-off for rejecting the null hypothesis is

* Correspondence: b.ylstra@vumc.nl

[1] Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands

Full list of author information is available at the end of the article

a delicate balance between sensitivity and specificity; while a stringent cut-off lowers the FDR, it also lowers average power and *vice* versa.The only way to improve both, or one without affecting the other, is to increase the number of biological replicates and thus perform more arrays. Sample size calculations can generally be divided into two categories. The first category asks the user to define values for certain parameters, such as the effect size (fold change of a differentially expressed gene) and the proportion of genes that are truly differentially expressed [5-9]. The second category estimates these parameters from existing data [10,11]. The method proposed here follows the latter approach and therefore requires pilot data.

To adapt mRNA expression array power calculations for aCGH and copy number changes, two key aspects need to be taken into account. Instead of concentrations of individual mRNA molecules, the underlying biology measured by aCGH consists of blocks of chromosomal DNA. Each block is (presumably) present in a normal copy number of two, but may contain areas of one or two-copy losses and one or more gains. Higher level amplifications can also be present. The aberrations contain both driver and passenger genes, and the breakpoints may vary from one sample to another.

As the entity being measured is DNA present in a discrete number of copies (0, 1, 2, 3, 4, ...), but individual array elements yield $\log_2$ ratios, aCGH data preprocessing generally involves the following steps that aim to better capture the biological relevance. *Normalization* first removes technical artifacts and makes the $\log_2$ ratios comparable across different hybridizations. *Segmentation* then identifies areas that share a common copy number and are separated by breakpoints. Finally, *calling* determines a discrete copy number level for each segment. At the moment, there is no clear consensus regarding the optimal stage of preprocessing from which the data should be used for downstream analysis. We discussed the topic and proposed that in most cases the recommended choice be to use calls, which have the clear advantage of having an attached biological meaning [12]. For power calculations however, the use of calls is problematic, as it would require the use of the chi-square test, for which no method of sample size calculation in large FDR-based multiple testing contexts is presently available. While both normalized and segmented log ratios allow the use of a t-test, they fail to take full advantage of the adjacency of consecutive array elements. Aberrations typically show great variation in their sizes ranging from focal amplifications to gains and losses of entire chromosome arms. Working directly with the original array elements does not take this into account, and gives larger aberrations significantly more weight than smaller ones as they contain more array elements. A possible improvement is therefore to replace array elements with regions, which are defined as a series of neighboring array elements sharing the same copy number signature. This reduces dimensionality with little loss of information [13]. Throughout this paper, the term *regions* is used to refer to the results of this analysis step.

For CGHpower, we are combining the advantages of regions with the feasibility of log ratios, by replacing the hard calls with median log ratios of all the array elements within a region. Together with these region-wise log ratios (RWLRs), the regions are then taken as a representation of the underlying biology (*i.e.* chromosomal regions with varying copy number levels). Each region is coupled to a null hypothesis stating that the means of the two groups do not differ from each other, which is the framework required for the power calculations proposed here. Regions that have a true difference between the two groups (generally normal copy number in one group and a gain, loss or amplification in the other) will be referred to as "differentially behaving regions".

After this preprocessing, power calculations are performed using regions as Ferreira *et al.* [14] previously described for both real and simulated gene expression data. T-statistics and p-values are calculated for each region from the RWLRs. All p-values from non-differentially behaving regions are expected to follow a uniform distribution, while those from the differentially behaving ones should follow another, unknown distribution (G). Two separate estimators of G are calculated: a non-parametric ($\tilde{G}_n$) and a parametric one ($\hat{G}_n$), which assumes that G follows a normal distribution. Both of these estimators depend on another unknown parameter, γ, which is the proportion of non-differentially behaving regions. When the estimate of γ used to calculate $\hat{G}_n$ and $\tilde{G}_n$ moves away from its true value, the difference between the two G estimators increases. The estimate of γ is therefore chosen so that this difference is minimized. The limiting density of effect sizes (λ) is then estimated using deconvolution, and so is G. Once these estimates have been calculated, approximate sample size calculations can be made using an adaptive version of the Benjamini-Hochberg method for multiple testing. While the original method [4] allows control over the FDR, the adaptive version also allows the estimation of average power [10].

While optimizing the protocol, there were certain options that we considered: whether to calculate the RWLRs as the mean or median of the log ratios, whether to use the Student's t-test assuming equal variances or Welch's t-test that allows unequal variances, and finally whether to calculate the p-values from normal or Student's t-distribution. All of the possible combinations were tested, and the optimum performance was observed with median log ratios, unequal variances and the normal

distribution. These choices were then fixed in CGHpower.

Implementation

Evaluation Data Sets

To evaluate the performance of CGHpower, eight recently published aCGH data sets that could be divided into two groups were collected. They will be referred to as Chin *et al.* [15], Douglas *et al.* [16], Fridlyand *et al.* [17], Myllykangas *et al.* [18], Nymark *et al.* [19], Postma *et al.* [20], Smeets *et al.* [21] and Wrage *et al.* [22] A total of five different array types were used among the data sets: VUmc 30 K spotted oligo [23] for data sets [15,20,22], Agilent Human 1 cDNA Microarray for [18,19], 3 K BAC array [24] for [16], 2 K BAC array [25] for [17] and 6 K BAC array for [21]. Table 1 provides a summary of the cancer and array types, together with group definitions and sizes.

Simulated Data Sets

In addition to real data sets, evaluation was also performed with simulated data. While generating the simulations, we attempted to implement realistic aspects of both signal and noise of tumor profiles. In the context of an aCGH experiment comparing two groups, the signal consists of aberrant regions that are specific to one of the groups. Noise consists of regions common to both groups, random aberrations in individual samples, and technical noise. Further characteristics are also that the sizes of the aberrant regions vary from entire chromosomes to focal aberrations, the exact start and end positions of a region vary slightly from one sample to another, and even a "common" region might not be be present in all of the samples.

The simulated data were generated by introducing artificial aberrations into a data set of clinical genetics samples of patients with mental retardation and no or few chromosomal aberrations [26]. To achieve a simulated data set of the desired size, resampling was performed with replacement. Aberrant regions were then randomly introduced as follows. A single array element was chosen at random as the starting point of a region. The size of the region was then chosen at random with a 10% probability for a single cytoband, 30% for three consecutive bands, 30% for six consecutive bands, 20% for the whole chromosome arm, and 10% for the entire chromosome. The type of the aberration was randomly chosen as a gain or loss with equal probabilities, but for the smallest aberrations of individual cytobands, a 2% probability for amplifications was also included. When introducing a region to a set of samples, the exact samples receiving the aberration were sampled from the Bernoulli distribution with p = 70%. Randomness was also introduced to the exact start and end positions of aberrations in individual samples by shifting the starting and ending array elements by a random number between -10 and 10.

A simulated data set of 15 + 15 arrays was generated with 30 common regions, and 5 regions for each individual sample. These copy number changes do not separate the two groups from each other, and therefore represent background noise. This data set is referred to as Simulation 0. Single regions specific to the two groups were then

Table 1: Evaluation data sets

Data Set	Array Type	Probes	Regions	Cancer Type	Groups (Samples)
Chin *et al.*	spotted oligo	26,755	223	breast	ER+ (113) *vs.* ER- (57)
Douglas *et al.*	BAC	3,032	142	colorectal	MSI (7) *vs.* CIN (30)
Fridlyand *et al.*	BAC	1,877	231	breast	TP53+ (10) *vs.* TP53- (52)
Myllykangas *et al.*	cDNA	11,342	260	gastric	diffuse (15) *vs.* intestinal (23)
Nymark *et al.*	cDNA	10,953	242	lung	asbestos-exposed (11) *vs.* non-exposed (9)
Postma *et al.*	spotted oligo	26,755	111	colorectal	good (16) *vs.* bad response (16)
Smeets *et al.*	BAC	4,196	143	head and neck	HPV+ (12) *vs.* HPV- (12)
Wrage *et al.*	spotted oligo	25,549	23	lung	BM+ (13) *vs.* BM- (15)
Simulation 0	in-situ oligo	42,331	440		(15) *vs.* (15)
Simulation 5	in-situ oligo	42,331	489		(15) *vs.* (15)
Simulation 10	in-situ oligo	42,331	525		(15) *vs.* (15)

Eight public data sets were collected to evaluate the performance of CGHpower. They represented five different cancer types and BAC, cDNA and oligo-based microarray platforms, with resolutions varying from 2 K to 27 K array elements. The last column contains the distinguishing factor used to divide the data set into two groups, along with the number of arrays in each group. The simulated data sets were generated by introducing artificial aberrations into a set of clinical genetics samples. A total of 11 simulations were generated, and the remaining ones are available at http://www.cangem.org/cghpower/. ER = estrogen receptor, MSI = microsatellite instability, CIN = chromosomal instability, HPV = human papilloma virus, BM = bone marrow metastasis.

introduced to Simulation 0 yielding data set Simulation 1. This process was repeated ten times resulting in a set of 11 simulations with the amount of differential signal ranging from none in Simulation 0 to 10 regions specific to each group in Simulation 10. Only Simulations 0, 5 and 10 are presented in this paper, but the full CGHpower outputs for all of them are available on the program's web page.

Preprocessing

All evaluation data sets were preprocessed starting from raw $\log_2$ ratios. First, the data were median normalized. Wavy patterns typically seen in many aCGH profiles were removed [26] from the 30 K arrays [15,20,22]. Normalized log ratios were segmented using the DNAcopy algorithm [27] and called by CGHcall [28] to identify gains, losses and amplifications. Regions between breakpoints were then collapsed into single data points, when shared between most of the samples [13]. Finally, the median log ratio was calculated for each of these regions in each sample, resulting in region-wise log ratios (RWLRs). All algorithms were run with default parameters, and sex chromosomes were excluded from the data.

Sample Size Calculations

For each region, t-statistics were calculated with a Welch's t-test allowing unequal variances and p-values computed from the normal distribution. The proportion of non-differentially behaving regions (γ) was estimated by minimizing the difference between parametric ($\hat{G}_n$) and non-parametric ($\tilde{G}_n$) estimators of *G*, which is the unknown distribution of the p-values from differentially behaving regions. The limiting density of effect sizes (λ)and *G* were then estimated using deconvolution. Finally, with FDR fixed at 10%, these parameter estimates were used to approximate average power as a function of sample size.

Results and Discussion

Estimates of average power as a function of sample size were calculated for the eight evaluation data sets and 11 simulations (Figure 1). The reliability of the power calculations depends directly on the the quality of parameter estimation, which in turns depends on compliance with required assumptions. The first assumption is that the proportion γ of non-differentially behaving regions be "substantially" smaller than 1 *(e.g.* ~0.9 will typically do, but 0.99 will not). The second assumption is that the RWLRs be approximately normally distributed, being neither particularly asymmetric (skewness) nor heavily tailed or extremely peaky (kurtosis). The complete CGHpower program output contains diagnostic plots from different stages of the power calculations procedure. These plots help determine to which extent these assumptions are fulfilled. While it is impossible to know what the true values of γ and λ are, one can easily evaluate how well the two estimators of *G* agree with each other (the "goodness-of-fit"). If they show a clear discrepancy, the accuracy of parameter estimation is questionable and the resulting power calculations consequently unreliable. Different scenarios in the quality of parameter estimation observed with the evaluation data sets are examined for each of the data sets to estimate the reliability of the calculated power.

The data sets Douglas *et al.*, Smeets *et al.*, Fridlyand *et al.* and Chin *et al.* are examples where the goodness-of-fit of the *G* estimators was satisfactory, ranked in this order according to their fits (Figure 2A). What appears to be the most important factor distinguishing these data sets from the others, is the density of the p-values. If there is no difference detected between two groups, p-values are expected to follow a uniform distribution, and their density function appears as a flat line. When the number of differentially behaving regions increases (γ moves away from 1), density at low p-values increases and the function is expected to be convex (Figure 2B). This can also be seen on the simulations where the amount of differential signal gradually increases from Simulation 0 to Simulation 10. Along with the increase in density for low p-values, also the goodness-of-fit systematically improves (data and figures at http://www.cangem.org/cghpower/).

Less satisfactory performance was observed with data sets of Postma *et al.* and Myllykangas *et al.* The goodness-of-fit shows more disagreement between the two estimators of *G* (Figure 2C) and as a result power estimates are less reliable. The density is increasing for low p-values, but slightly less and the function is not convex as expected (Figure 2D). Compared to Simulation 0, which has no true differences between the groups, the increase in p-value density for the data set of Myllykangas *et al.* is very small. One explanation is that there is simply not enough differential signal that is detectable with a t-test. Alternatively, the number of differentially behaving regions might be too low (*i.e.* γ is too close to 1). While these data sets do give γ estimates of 0.75 and 0.55, respectively, these estimates cannot be trusted if the estimates of *G* disagree with each other. Therefore it is recommended that the goodness-of-fit plot be used to assess the reliability of the estimates of other parameters. Also, judging from the results with the simulated data sets, CGHpower seems to underestimate the true value of γ.

While assumptions regarding γ seem to be most important, the RWLRs are also assumed to be normally distributed. The program output contains histograms of the skewness (asymmetry) and kurtosis (peakedness) of the RWLRs, superimposed with those of a normal distribution (data on the CGHpower web page). Assumptions of

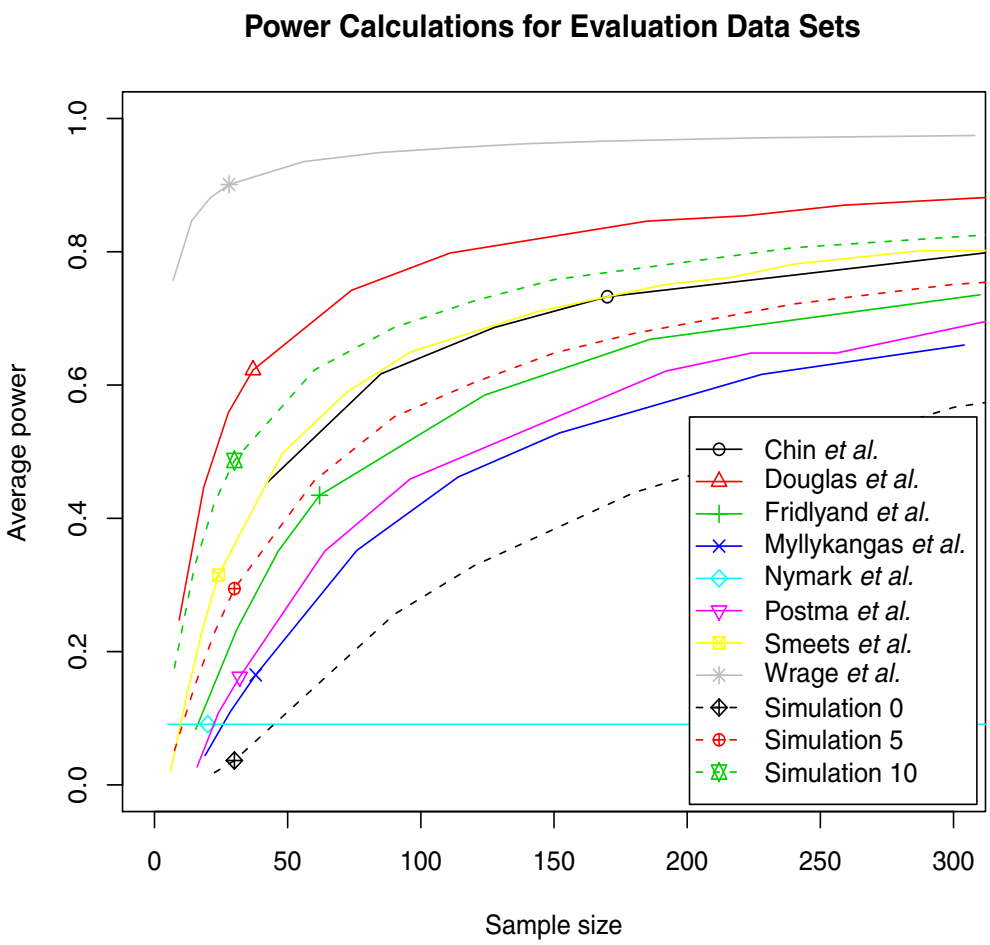

Figure 1 Power calculations for evaluation data sets. Average power estimated as a function of sample size for the eight evaluation data sets and three simulations. False discovery rate was fixed at 10%. The horizontal position of the small symbols mark the actual size of the data set that was used to calculate the estimates in each case. Real data sets are shown with solid lines and three of the simulations with dotted lines. Additional simulations are available at http://www.cangem.org/cghpower/.

normality become more critical with small sample sizes and less important with large ones. Within the evaluation data sets, most violations of normality were observed with the Chin *et al.* data set, yet this is one of the better-performing ones in terms of goodness-of-fit. This might be explained by the relatively large sample size (170) of the study. Another factor besides the number of arrays, is the number of regions found after the preprocessing step. The larger the number of regions, the better the performance of the parameter estimation and therefore the reliability of power calculations. The assumption of normality is therefore more crucial with samples containing very few biological differences.

The data sets of Nymark *et al.* and Wrage *et al.* are examples where our method failed to work, despite the differences reported and technically as well as biologically validated. In the case of Nymark *et al.* the obtained power curve is a flat line (Figure 1). This can happen when parameter estimation fails. The explanation can be found from the density of the p-values, but now the assumptions were violated more severely than in the cases of Postma *et al.* and Myllykangas *et al.* The density function is actually concave and shows even less density at low p-values than would be expected by chance (Figure 2F). With Wrage *et al.*, failure can be observed at the preprocessing step, as only 23 regions are detected (Table 1).

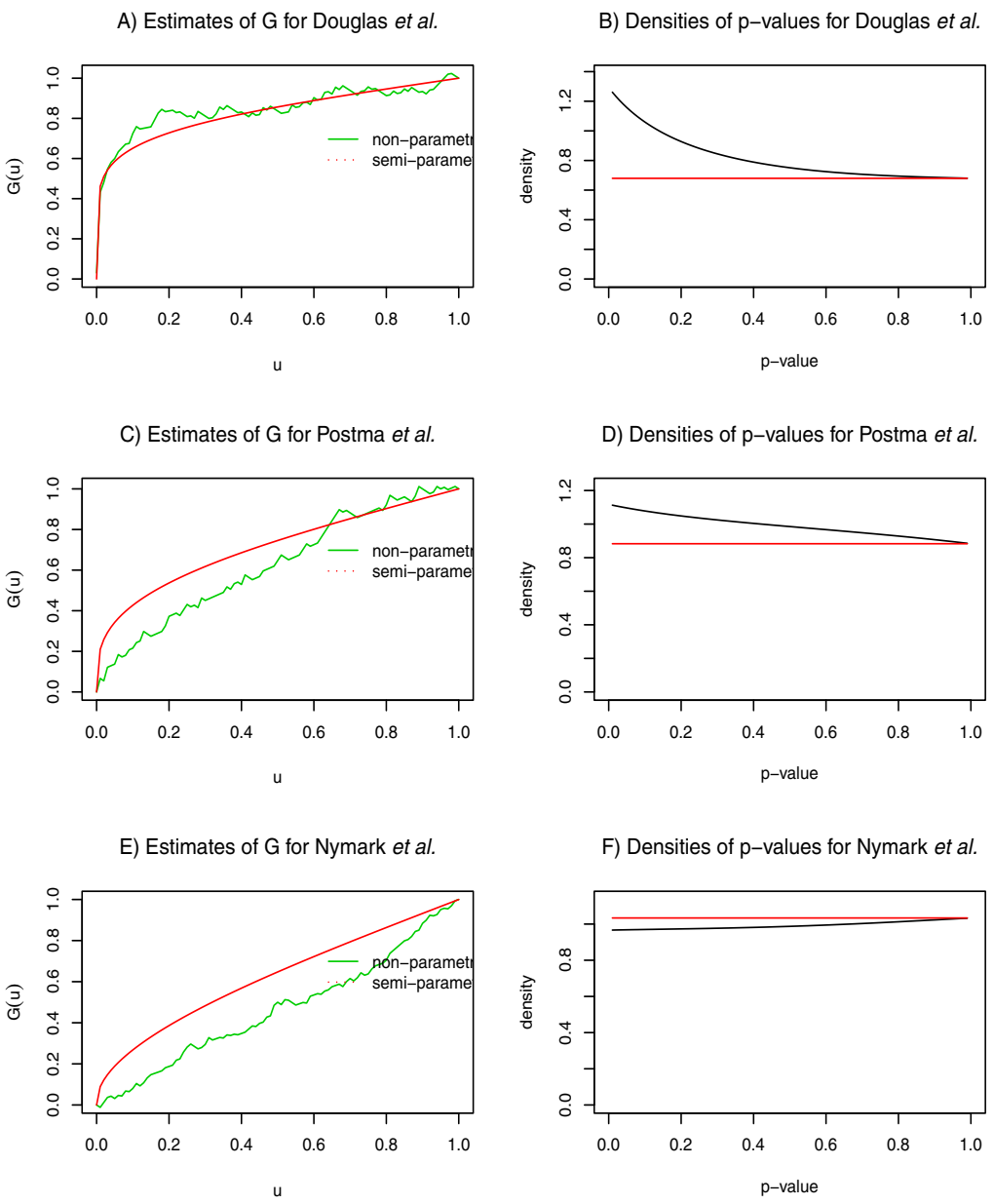

Figure 2 Diagnostic plots. The goodness-of-fit of the two estimators of *G* and densities of p-values for three data sets illustrating different scenarios in the performance of CGHpower. The data set of Douglas *et al.* shows A) a satisfactory goodness-of-fit following from B) a convex p-value density function. Mediocre operation is demonstrated with the data set of Postma *et al.* C) An inferior fit results from D) a p-value density which shows a slight increase for small values, but is not convex as expected. Nymark *et al.* represents failed execution. E) The disagreement between the *G* estimators is slightly more severe and the estimated power curve is a flat line (Figure 1). F) P-values exhibit even less density at low values than would be expected by chance. In such circumstances, it is recommended that data preprocessing be carried out before uploading and only the power calculations part be performed in CGHpower.

Since the sex chromosomes are excluded from the analysis, this means that only one copy number breakpoint was detected in the whole genome using the fixed CGHpower preprocessing described above. As preprocessing and power calculations procedures are fixed earlier in CGHpower, it was not optimized it for every aCGH platform or data set. Allowing the user to fine-tune different settings and immediately see the result of each change would require implementing a more complex user interface, similar to desktop software, which would be impractical for a single-purpose web tool. As an alternative option, if the goodness-of-fit and density plots indicate that power calculations failed, users can perform preprocessing independently, turn off the preprocessing step from the program, and perform the power calculations only.

Consistency as the Pilot Size Is Increased

CGHpower was initially developed to be used on smaller pilot data sets in the planning stages of larger microarray experiments or for verifying power achieved in past experiments. We wanted to evaluate whether the resulting power estimates hold while more and more arrays are added to the data set. Assuming that a pilot of 10 + 10 arrays has estimated an experiment with 40 + 40 arrays should result in an average power of approximately 70%. The data set of 80 arrays is then generated and for verification the power calculations are repeated with the entire data set. If the new results indicate that the achieved power is in fact only 50%, and that 20 + 20 new arrays are needed in order to achieve our goal of 70%, then the two power calculations have to be declared inconsistent. To evaluate whether the power estimates remain consistent while the pilot size is increased, power was calculated with smaller subsets of the Chin *et al.* data set, since it is our largest one. This data set contains a total of 170 arrays (113 *vs.* 57), which was split into smaller subsets to represent pilots of a larger study. Nine resamplings ranging from 10% (11 *vs.* 6 arrays) to 90% (102 *vs.* 51) of the original data set were randomly selected for the power calculations. Each resampling was repeated 10 times and the results were averaged. Two of the ten repetitions of the 10% subset and one repetition in the 20% subset experienced a failed power estimation resulting in flat power curves as with the Nymark *et al.* data set. These cases were removed before averaging the results. A plot of the resulting power estimates shows that except for the smallest subset (11 *vs.* 6 arrays), the results appear to be consistent (Figure 3). This suggests that as long as the pilot is of sufficient size, power estimates generated with CGHpower using smaller pilot data sets are in fact representative of a subsequent larger study. While the exact requirement for a "sufficient pilot" is hard to define beforehand, the power calculations can be repeated when more arrays are performed to see whether power estimates are still changing or have been stabilized.

Conclusions

We have explored sample size calculations in the context of aCGH and copy number changes and propose a dedicated tool for this purpose. From a pilot data set, CGHpower estimates the biological diversity between two groups of cancer samples and estimates average power as a function of sample size using an adaptive version of the Benjamini-Hochberg method for multiple testing [4,10]. Pilot data is used for parameter estimation and this requires certain assumptions to hold in an approximate sense. We have evaluated the performance of CGHpower with eight published data sets, four of which show satisfactory performance using predefined preprocessing measures. Among these data sets were BAC and oligo-based array platforms, whose resolution varied from less than 2 K for BACs to almost 27 K for oligos. The differences in resolution did not have a direct impact on the obtained power estimates, which should be determined more by the amount of biological variation between the two groups.

In two data sets violations of critical assumptions lead to problems in parameter estimation and therefore power estimates are less reliable. More severe violations and/or the inflexibility of a completely predefined analysis procedure lead to failed execution for the two other data sets. Even though the proposed method has its limitations, it is to our knowledge the only proposed one for aCGH data and copy number changes. As the program allows performance evaluation through diagnostic plots, critical judgement can be applied for each data set.

As a summary on the evaluation of CGHpower results, users should consider paying attention to the following: 1) Do the copy number profile plots appear similar to the aberrations that you have detected in your own analysis? If CGHpower does not seem to detect the important aberrations, consider performing the preprocessing before uploading and use CGHpower only for the power calculations. 2) Do the estimators of G agree with each other? If the goodness-of-fit is poor, so will other parameter (and resulting power) estimates. 3) Is the density function of the p-values convex, and showing a higher density at small p-values? A straight or concave function might be caused by too small effect size, or γ being too close to one. 4) Excess skewness and/or kurtosis in the data might also affect the performance, but this seems to be less crucial.

The proposed method uses log ratios instead of calls, even though we feel the latter is generally the preferred choice when working with aCGH data. Calls have the benefit of a clear biological meaning and are therefore easier to interpret. However, their use for power calcula-

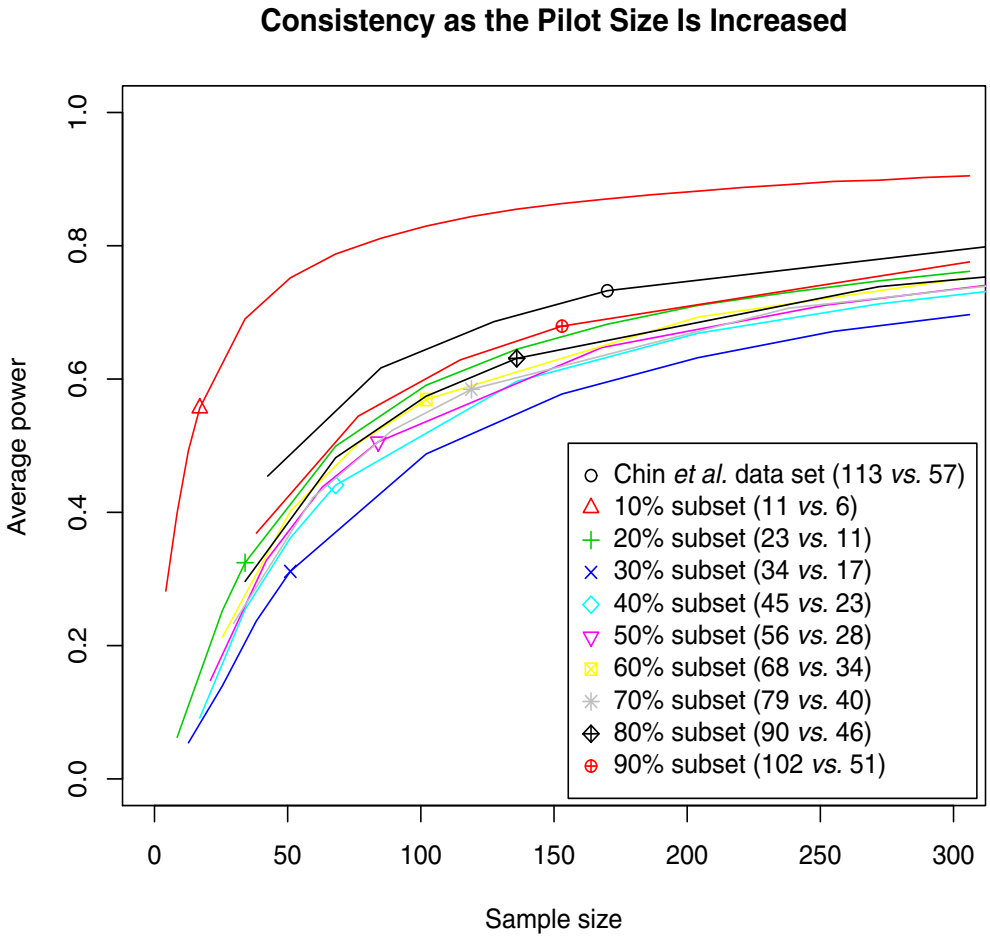

Figure 3 Consistency as the pilot size is increased. To evaluate whether power estimates obtained from smaller pilots are in fact representative of larger data sets, the calculations were performed with subsets of the Chin *et al.* data. Resampling without replacement was used to obtain subsets from 10% to 90% of the original data set. Each resampling was repeated ten times and results averaged. The horizontal position of the small symbols mark the size of the subset used to obtain each power estimate.

tions in the context of FDR is problematic, as it would require using the chi-square test, a setting that is not as well developed as the Gaussian one. Also, as log ratios are the basis for calls in the first place, they do contain all the necessary information even though they are not as clear to interpret.

In comparison to sample size calculations for mRNA expression arrays, the differentiating factor for aCGH studies is the concept of regions, which stems from the different biological phenomenon underlying the microarray $\log_2$ ratios. Compared to the number of array elements, the number of regions is relatively small, which presents challenges to parameter estimation from the data. As the total number of regions is remarkably smaller than with expression arrays, the estimation might fail if the number of differentially behaving regions is too small, even if there is a true difference between the groups.

An important concern when performing power calculations is the actual power requirement. A power curve typically plateaus out at some point, indicating saturation. Increasing the average power from *e.g.* 60% to 70% requires a significantly bigger increase in sample size than is needed for an increase from 50% to 60%. Therefore it is

difficult to set a a predefined gold standard of adequate power. One option is to try to find where the slope of the power curve is decreasing rapidly. This should give a reasonable compromise between statistical power and cost of the experiment. Another aspect worth pointing out, is that the level of power needed also depends on the research question. For example, if the goal is to construct a classifier that can classify future samples into one of the two groups, a lower level of average power might yield a perfectly satisfactory classifier even though not all differences are detected.

Availability and requirements

CGHpower is a web-based application and can be freely accessed at http://www.cangem.org/cghpower/. It allows direct uploads and can also automatically retrieve data stored in the CanGEM database [29]. The computation times of CGHpower may vary considerably depending on the number of samples and array elements in the data set, and also on the prevailing load of the Linux cluster where the calculations are performed. As an example, running times for a data set of 30 samples and 42 K array elements have been around 1-1.5 hours in our test runs. The software has been implemented in R [30] and the source code is available upon request.

Authors' contributions

BY conceived the study. IS, JAF, MAW and BY designed CGHpower. IS performed the implementation. IS, JAF, MAW and BY wrote the manuscript with critical comments from SK and GAM. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Finnish special governmental subsidies for health sciences research in Helsinki University Central Hospital; the Finnish Funding Agency for Technology and Innovation (TEKES, 40141/07); the Sigrid Jusélius Foundation; the Centre for Medical Systems Biology (CMSB); and the Centre of Excellence Approved by the Netherlands Genomics Initiative/Netherlands Organisation of Scientific Research (NWO); and this study was performed within the framework of CTMM, the Center for Translational Molecular Medicine. DeCoDe project (grant 03O-101).

FIMM Technology Centre, Institute for Molecular Medicine Finland (FIMM) is acknowledged for their high performance computing facilities.

Author Details

[1]Department of Pathology, VU University Medical Center, Amsterdam, The Netherlands, [2]Department of Pathology, Haartman Institute and HUSLAB, University of Helsinki and Helsinki University Central Hospital, Finland, [3]FIMM Technology Centre, Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland, [4]Department of Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands and [5]Department of Mathematics, Vrije Universiteit, Amsterdam, The Netherlands

Received: 14 December 2009 Accepted: 17 June 2010
Published: 17 June 2010

References

1. Kallioniemi A: **CGH microarrays and cancer.** *Curr Opin Biotechnol* 2008, **19:**36-40.
2. Shinawi M, Cheung SW: **The array CGH and its clinical applications.** *Drug Discov Today* 2008, **13(17-18):**760-770.
3. van de Wiel MA, Smeets SJ, Brakenhoff RH, Ylstra B: **CGHMultiArray: exact P-values for multi-array comparative genomic hybridization data.** *Bioinformatics* 2005, **21(14):**3193-3194.
4. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Stat Soc Ser B* 1995, **57:**289-300.
5. Lee MLT, Whitmore GA: **Power and sample size for DNA microarray studies.** *Stat Med* 2002, **21(23):**3543-3570.
6. Muller P, Parmigiani G, Robert C, Rousseau J: **Optimal sample size for multiple testing: the case of gene expression microarrays.** *J Am Stat Assoc* 2004, **99(468):**990-1001.
7. Pan W, Lin J, Le CT: **How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach.** *Genome Biol* 2002, **3(5):**research 0022.
8. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A: **False discovery rate, sensitivity and sample size for microarray studies.** *Bioinformatics* 2005, **21(13):**3017-3024.
9. Tibshirani R: **A simple method for assessing sample sizes in microarray experiments.** *BMC Bioinformatics* 2006, **7:**106.
10. Ferreira JA, Zwinderman AH: **Approximate power and sample size calculations with the Benjamini-Hochberg method.** *Int J Biostat* 2006, **2(10):**Article 8.
11. Jørstad TS, Midelfart H, Bones AM: **A mixture model approach to sample size estimation in two-sample comparative microarray experiments.** *BMC Bioinformatics* 2008, **9:**117.
12. van de Wiel MA, Picard F, van Wieringen WN, Ylstra B: **Preprocessing and downstream analysis of microarray DNA copy number profiles.** *Brief Bioin-form* 2010 in press.
13. van de Wiel MA, van Wieringen WN: **CGHregions: dimension reduction for array CGH data with minimal information loss.** *Cancer Informatics* 2007, **3:**55-63.
14. Ferreira JA, Zwinderman A: **Approximate sample size calculations with microarray data: an illustration.** *Stat Appl Genet Mol Biol* 2006, **5(1):**Article 25.
15. Chin SF, Teschendorff AE, Marioni JC, Wang Y, Barbosa-Morais NL, Thorne NP, Costa JL, Pinder SE, van de Wiel MA, Green AR, Ellis IO, Porter PL, Tavare S, Brenton JD, Ylstra B, Caldas C: **High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer.** *Genome Biol* 2007, **8(10):**R215.
16. Douglas EJ, Fiegler H, Rowan A, Halford S, Bicknell DC, Bodmer W, Tomlinson IPM, Carter NP: **Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas.** *Cancer Res* 2004, **64(14):**4817-4825.
17. Fridlyand J, Snijders AM, Ylstra B, Li H, Olshen A, Seg-raves R, Dairkee S, Tokuyasu T, Ljung BM, Jain AN, McLennan J, Ziegler J, Chin K, Devries S, Feiler H, Gray JW, Waldman F, Pinkel D, Albertson DG: **Breast tumor copy number aberration phenotypes and ge-nomic instability.** *BMC Cancer* 2006, **6:**96.
18. Myllykangas S, Junnila S, Kokkola A, Autio R, Scheinin I, Kiviluoto T, Karjalainen-Lindsberg M, Hollmen J, Knuu-tila S, Puolakkainen P, Monni O: **Integrated gene copy number and expression microarray analysis of gastric cancer highlights potential target genes.** *Int J Cancer* 2008, **123(4):**817-825.
19. Nymark P, Wikman H, Ruosaari S, Hollmen J, Vanhala E, Karjalainen A, Anttila S, Knuutila S: **Identification of specific gene copy number changes in asbestos-related lung cancer.** *Cancer Res* 2006, **66(11):**5737-5743.
20. Postma C, Koopman M, Buffart TE, Eijk PP, Carvalho B, Peters GJ, Ylstra B, van Krieken JH, Punt CJA, Meijer GA: **DNA copy number profiles of primary tumors as predictors of response to chemotherapy in advanced colorectal cancer.** *Ann Oncol* 2009, **20(6):**1048-1056.
21. Smeets SJ, Braakhuis BJM, Abbas S, Snijders PJF, Ylstra B, van de Wiel MA, Meijer GA, Leemans CR, Brak-enhoff RH: **Genome-wide DNA copy number alterations in head and neck squamous cell carcinomas with or without oncogene-expressing human pa-pillomavirus.** *Oncogene* 2006, **25(17):**2558-2564.
22. Wrage M, Ruosaari S, Eijk PP, Kaifi JT, Hollmen J, Yekebas EF, Izbicki JR, Brakenhoff RH, Streichert T, Riethdorf S, Glatzel M, Ylstra B, Pantel K, Wikman H: **Genomic profiles associated with early micrometas-tasis in lung cancer: relevance of 4q deletion.** *Clin Cancer Res* 2009, **15(5):**1566-1574.

23. van den Ijssel P, Tijssen M, Chin SF, Eijk P, Carvalho B, Hopmans E, Holstege H, Bangarusamy DK, Jonkers J, Meijer GA, Caldas C, Ylstra B: **Human and mouse oligonucleotide-based array CGH.** *Nucleic Acids Res* 2005, **33(22):**e192.
24. Fiegler H, Carr P, Douglas EJ, Burford DC, Hunt S, Scott CE, Smith J, Vetrie D, Gorman P, Tomlinson IPM, Carter NP: **DNA microarrays for comparative ge-nomic hybridization based on DOP-PCR amplification of BAC and PAC clones.** *Genes Chromosomes Cancer* 2003, **36(4):**361-74.
25. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG: **Assembly of microarrays for genome-wide measurement of DNA copy number.** *Nat Genet* 2001, **29(3):**263-264.
26. van de Wiel MA, Brosens R, Eilers PHC, Kumps C, Meijer GA, Menten B, Sistermans E, Speleman F, Timmerman ME, Ylstra B: **Smoothing waves in array CGH tumor profiles.** *Bioinformatics* 2009, **25(9):**1099-1104.
27. Venkatraman ES, Olshen AB: **A faster circular binary segmentation algorithm for the analysis of array CGH data.** *Bioinformatics* 2007, **23(6):**657-663.
28. van de Wiel MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, Ylstra B: **CGHcall: calling aberrations for array CGH tumor profiles.** *Bioinformatics* 2007, **23(7):**892-894.
29. Scheinin I, Myllykangas S, Borze I, Bohling T, Knuutila S, Saharinen J: **CanGEM: mining gene copy number changes in cancer.** *Nucleic Acids Res* 2008, **36(Database):**D830-D835.
30. R Development Core Team: **R: A Language and Environment for Statistical Computing.** 2009 [http://www.R-project.org]. R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0

doi: 10.1186/1471-2105-11-331
Cite this article as: Scheinin *et al.*, CGHpower: exploring sample size calculations for chromosomal copy number experiments *BMC Bioinformatics* 2010, **11**:331

Chapter 4

DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly

Ilari Scheinin[1]
Daoud Sie[1]
Henrik Bengtsson
Mark A van de Wiel
Adam B Olshen
Hinke F van Thuijl
Hendrik F van Essen
Paul P Eijk
François Rustenburg
Gerrit A Meijer
Jaap C Reijnveld
Pieter Wesseling
Daniel Pinkel
Donna G Albertson
Bauke Ylstra

[1] Shared first authors.

This publication has also been included in Daoud Sie's dissertation,
Breaking the Cancer Genome Code for Patient Care,
VU University Amsterdam, 2017.

Genome Research (2014) **24:** 2022–2032

DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly

Ilari Scheinin,[1,2,12] Daoud Sie,[1,12] Henrik Bengtsson,[3,4] Mark A. van de Wiel,[5,6] Adam B. Olshen,[3,4] Hinke F. van Thuijl,[1,7] Hendrik F. van Essen,[1] Paul P. Eijk,[1] François Rustenburg,[1] Gerrit A. Meijer,[1] Jaap C. Reijneveld,[7,8] Pieter Wesseling,[1,9] Daniel Pinkel,[3,10] Donna G. Albertson,[3,10,11] and Bauke Ylstra[1]

[1]*Department of Pathology, VU University Medical Center, 1007 MB Amsterdam, The Netherlands;* [2]*Department of Pathology, Haartman Institute and HUSLAB, FIN-00014 University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland;* [3]*Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, California 94158, USA;* [4]*Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California 94158, USA;* [5]*Department of Epidemiology and Biostatistics, VU University Medical Center, 1007 MB Amsterdam, The Netherlands;* [6]*Department of Mathematics, VU University, 1181 HV Amsterdam, The Netherlands;* [7]*Department of Neurology, VU University Medical Center, 1007 MB Amsterdam, The Netherlands;* [8]*Department of Neurology, Academic Medical Centre, 1105 AZ Amsterdam, The Netherlands;* [9]*Department of Pathology, Radboud University Medical Centre, 6500 HB Nijmegen, The Netherlands;* [10]*Department of Laboratory Medicine, University of California San Francisco, San Francisco, California 94153, USA;* [11]*Bluestone Center for Clinical Research, New York University College of Dentistry, New York, New York 10010-4086, USA*

Detection of DNA copy number aberrations by shallow whole-genome sequencing (WGS) faces many challenges, including lack of completion and errors in the human reference genome, repetitive sequences, polymorphisms, variable sample quality, and biases in the sequencing procedures. Formalin-fixed paraffin-embedded (FFPE) archival material, the analysis of which is important for studies of cancer, presents particular analytical difficulties due to degradation of the DNA and frequent lack of matched reference samples. We present a robust, cost-effective WGS method for DNA copy number analysis that addresses these challenges more successfully than currently available procedures. In practice, very useful profiles can be obtained with ~0.1× genome coverage. We improve on previous methods by first implementing a combined correction for sequence mappability and GC content, and second, by applying this procedure to sequence data from the 1000 Genomes Project in order to develop a blacklist of problematic genome regions. A small subset of these blacklisted regions was previously identified by ENCODE, but the vast majority are novel unappreciated problematic regions. Our procedures are implemented in a pipeline called QDNAseq. We have analyzed over 1000 samples, most of which were obtained from the fixed tissue archives of more than 25 institutions. We demonstrate that for most samples our sequencing and analysis procedures yield genome profiles with noise levels near the statistical limit imposed by read counting. The described procedures also provide better correction of artifacts introduced by low DNA quality than prior approaches and better copy number data than high-resolution microarrays at a substantially lower cost.

[Supplemental material is available for this article.]

Alteration in chromosomal copy number is one of the main mechanisms by which cancerous cells acquire their hallmark characteristics (Pinkel et al. 1998; Hanahan and Weinberg 2011). For > 20 yr, these alterations have been routinely detected first by genome-wide comparative genomic hybridization (CGH) (Kallioniemi et al. 1992) and subsequently by array-based CGH (Snijders et al. 2001) or single nucleotide polymorphism (SNP) arrays (Ylstra et al. 2006). Now whole-genome sequencing (WGS) offers an alternative to microarrays for many genome analysis applications, including copy number detection.

Several methods have been developed to estimate DNA copy number from WGS data. They can be grouped into the following four categories, each of which has its own set of requirements, strengths, and weaknesses (Teo et al. 2012): (1) Assembly-based methods construct the genome piece by piece from the sequence reads instead of aligning them to a known reference; these methods have the greatest sensitivity to detect deviations from the reference genome, including copy number changes and genome

[12]**These authors contributed equally to this work.**
Corresponding author: B.Ylstra@vumc.nl
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.175141.114. Freely available online through the *Genome Research* Open Access option.

24:2022–2032 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/14; www.genome.org

rearrangements, but require high sequence coverage (typically 40×) (Li et al. 2010) and therefore incur high cost; (2) split-read and (3) read-pair methods map sequence reads from both ends of size-fractionated genomic DNA molecules onto the reference genome; these methods can provide information on copy number and genome rearrangements, but they impose requirements on molecule sizes and therefore are highly sensitive to DNA integrity; and (4) depth of coverage (DOC) methods infer copy number from the observed sequence depth across the genome and do not require both ends of the molecule to be sequenced.

Archival tissue is an invaluable resource for biomarker detection studies (Casparie et al. 2007). Projects investigating cancers with long survival, such as diffuse low-grade gliomas (LGGs) with a subset of patients surviving > 25 yr after diagnosis (van Thuijl et al. 2012), require long-term clinical follow-up. Archival FFPE tissue is often the only source of material for study (Blow 2007). The use of such samples has been challenging due to poor DNA quality; hence, array CGH results, for example, have been variable (Mc Sherry et al. 2007; Hostetter et al. 2010; Krijgsman et al. 2012; Warren et al. 2012). To make large archival sample series accessible for genome research, a robust technique is required that performs well on diverse sample types, with high resolution, quality and reproducibility, and at low cost without the necessity for a (matched) normal sample. Here we focus exclusively on DOC methods, because they are theoretically most compatible with DNA isolated from FFPE material.

Typically, DOC methods for copy number divide the reference genome into bins and count the number of reads in each, although there are also bin-free intensity-based implementations (Shen and Zhang 2012). Copy number is then inferred from the observed read counts across the genome. To compensate for technological bias, many DOC algorithms, such as CNV-seq (Xie and Tammi 2009), SegSeq (Chiang et al. 2009), BIC-seq (Xi et al. 2011), and CNAnorm (Gusnanto et al. 2012), compare tumor signal to a normal reference signal, similar to array CGH. Commonly, a pool of different individuals is used as a normal reference DNA. In many applications, including cancer genome analysis, matched normal DNA from the same patient is preferable to avoid detection of germline copy number variants (Feuk et al. 2006), allowing focus solely on somatic aberrations (Perry et al. 2008).

Two DOC methods, readDepth (Miller et al. 2011) and FREEC (Boeva et al. 2011), do not require a reference signal. This has three principal advantages: the cost is reduced by half, archival material for which matched normal reference tissue is unavailable (most cases) can be analyzed, and measurement noise from the reference sample is avoided. Achieving these benefits requires accurate computational correction for biases in the DOC sequence data since they are no longer being normalized by comparison with data from a matched reference specimen.

Here we describe a multiplexed, single-read (SR), shallow WGS procedure based on the Illumina platform that produces improved DOC copy number profiles. Because DOC profiles are fundamentally based on counting the number of sequence reads, the minimum achievable noise can be easily calculated. We show that a larger proportion (most) of the samples we have analyzed with our procedures show noise levels at the theoretical minimum than with other analysis methods. We achieve the improved performance by simultaneous (rather than sequential) correction of primary read counts for sequence mappability and GC content, and by using a comprehensive empirical approach for recognition and filtering of problematic genome regions. We also show that compared to previous shallow WGS analysis procedures, our approach provides improved correction of spurious localized profile variations, which are presumably due to sample quality problems; and microarray analysis costs more and yields a poorer signal-to-noise ratio than shallow WGS. Thus our DOC profiles provide a more accurate representation of the genome copy number structure than can be obtained by other approaches and should allow segmentation and calling algorithms to more sensitively recognize true aberrations.

Results

Shallow WGS and alignment to the reference genome

Shallow WGS was performed with DNA isolated from FFPE sections of 15 LGGs (van Thuijl et al. 2014), two oral squamous cell carcinomas (SCCs AB042 and AB052) (Bhattacharya et al. 2011), and the breast cancer cell line BT474 on the Illumina HiSeq 2000 using run mode SR50, which sequences only one end of the DNA molecules for 50 base pairs. In general, these DNA samples were multiplexed with others so that each HiSeq sequencing lane contained between 18 and 22 total samples. We use sample LGG150 to illustrate our analysis procedures in the main article text and figures because it contains a range of different types of genome alterations that are typical for solid tumors. Complete analyses of all LGG samples, BT474, AB042, and AB052, including whole-genome plots and enlarged views of chromosome 1, are presented in Supplemental Figures S1–S3. In addition, we present noise data from more than 1000 mostly formalin-fixed archival specimens obtained from many hospitals throughout Europe.

On average, we obtained 9.2 million total reads per sample (range 3.1–23.9) for the multiplexed samples, of which 8.2 million (range 3.0–22.9) aligned to the human reference genome with the sequence alignment algorithm BWA (Li and Durbin 2009). We filtered out PCR duplicate reads and reads with mapping qualities lower than 37 (highest value returned by BWA), resulting in a final average read count of 6.0 million (range 2.4–18.1) per sample. Read counts for the 15 LGGs, AB042, AB052, and BT474 are provided in Supplemental Table S1.

Binning of sequence reads

We divided the human reference genome into nonoverlapping, fixed-sized bins. We use 15-kb bins in the analysis presented here because this results in approximately the same number of bins as the number of array elements on 180K oligonucleotide CGH arrays and provides reasonable noise levels with as few as 6 million reads. We note, however, that any bin size could be used, and such an option is provided in the accompanying software package, QDNAseq. Removal of 12,893 bins that were completely composed of uncharacterized bases (denoted with N's in the human reference genome sequence) resulted in a total number of 179,187 autosomal bins. We determined raw copy number estimates by counting the number of reads in each bin. The median-normalized $\log_2$-transformed read counts, the raw copy number profile, for sample LGG150 is shown in Figure 1A. Regions of low-level loss and gain (e.g., on chromosomes 10 and 20, respectively) are apparent in the profile. In addition, some very narrow regions of highly elevated read counts and a substantial number of bins with very low read counts are present. The horizontal stripes of data points are due to the integer nature of the read counts. Experience based on classical cytogenetics and array CGH suggests that many features of this profile reflect characteristics of the sequencing and analysis process rather than true copy number variation (Baldwin et al. 2008).

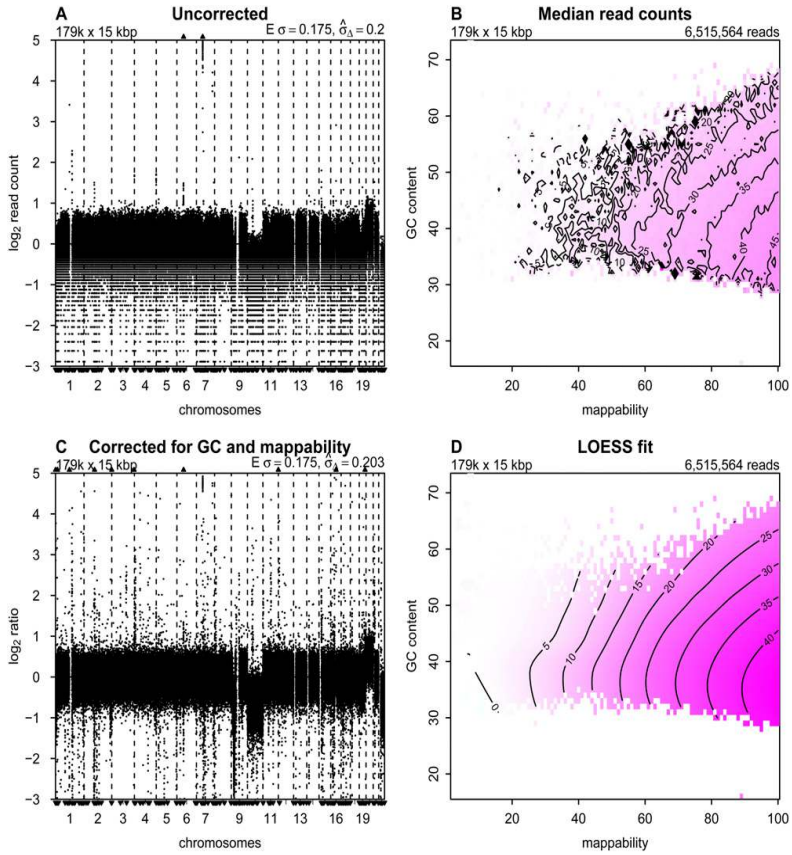

Figure 1. Correction to read counts. Copy number profiles from (*A*) uncorrected and (*C*) corrected read counts; (*B*) median read counts per bin as a function of GC content and mappability; and (*D*) the corresponding LOESS fit for sample LGG150. Regions of the isobar plots that are white contain no bins with that combination of GC and mappability. In the copy number profiles, bins are ordered along the *x*-axis by their genomic positions, and the *y*-axis shows median-normalized $\log_2$-transformed data. Small triangles at the *top* and *bottom* edges represent data points that fall outside the plot area. *Upper left* corners show the number and size of bins. *Upper right* corners of the median read counts plot shows the total number of sequence reads, and *upper right* corners of the copy number profiles the expected and measured standard deviation. The expected standard deviation (E σ) is defined as $\sqrt{1/N}$, where N is the average number of reads per bin. The measured standard deviation ($\hat{\sigma}_\Delta$) is calculated from the data with a mean-scaled and 0.1%-trimmed first-order estimate, prior to $\log_2$ transforming the data for plotting (see text).

Correction of read counts

It is well established that raw read counts are affected by GC content and mappability of the sequence reads (Benjamini and Speed 2012; Derrien et al. 2012; Rieber et al. 2013). Published analysis methods generally correct for these factors independently if corrections for both are used. Although independent correction is effective for many cases, genome profiles from some samples, especially those that are formalin-fixed, contain clearly artifactual variations. Independent correction for GC and mappability is appropriate only if these two factors do not interact in their effects on read counts. We desired to determine if simultaneous correction might provide improved read count profiles. We implemented simultaneous correction by calculating the median read count for all bins with the same combinations of GC and mappability (Fig. 1B). We then fit a LOESS surface through the medians (Fig. 1D). To correct the raw read count of a bin, we divided the raw count by the LOESS value of its combination of GC and mappability. (Fitting the LOESS has the benefit of stabilizing the values for bins with closely related GC and mappability.) Following this procedure, the corrected profile for LGG150, after $\log_2$-transformation and centering, is much cleaner (Fig. 1C) than before correction. The correction of bins with low counts is particularly noticeable, but at the cost of introducing bins with read counts that appear to be anomalously high. Copy number profiles and plots of the median read counts as a function of GC content and mappability are shown for the 15 LGGs, AB042, AB052, and BT474 in Supplemental Figures S1 and S2.

Blacklisting bins to exclude problematic regions

Examination of Figure 1C shows the presence of multiple very narrow peaks and some apparent deletions that might indicate aberrations. Some of these structures, for example many of the narrow peaks, appear to have been introduced by the GC-mappability correction. Many of these features are highly recurrent across, both tumor and normal, samples (data not shown). Recurrence alone may imply that these peaks represent common germline copy number variations (CNVs). The observation that they are frequently located in (peri-)centromeric and (sub-)telomeric regions, however, suggested that a large number are artifacts.

The presence of chromosomal regions with anomalous behavior is well established and has led others, for example, the ENCODE Project Consortium, to develop blacklists of sequences to exclude from their analyses (The ENCODE Project Consortium 2012). Some of these sequences map to regions with known repeat elements, such as satellites, centromeric, and telomeric repeats. Therefore we tested the effect of removing bins with mappabilities below the arbitrary threshold of 50 and bins overlapping with the ENCODE blacklists (Fig. 2A). Clearly, the profiles are improved, but many regions of potentially artifactual variation remain (indicated by black dots in Fig. 2A). Changing the mappability threshold affects the results to some degree but fails to sufficiently remove the problematic regions without also removing a major proportion of the bins (see Supplemental Fig. S4).

Given the insufficiency of the ENCODE blacklist for copy number analysis and the apparent recurrence of the problematic regions, we developed our own data-driven list of problematic genome regions. We started by analysis of a collection of normal genomes, which has the potential to identify problematic sequence motifs as in ENCODE, unknown problems in the reference genome sequence, and common CNVs. We obtained the required sequence data from the publicly available WGS data set from the 1000 Genomes Project (1000G) (The 1000 Genomes Project Consortium 2012). After selecting samples that were sequenced in a manner similar to our experimental setup (Illumina platform, low-coverage, SR50), we identified and downloaded 38 cases. The individuals have a substantial range of ethnic backgrounds (nine CEU, eight JPT, seven YRI, five CHB, three ASW, two PUR, one CLM, one IBS, one LWK, and one MXL).

The 38 samples were then processed as described above. The difference between the actual count and the LOESS fitted value was determined for each bin based on its GC and mappability values. These residuals were recorded for each sample, and the median of the residuals across the 38 samples was calculated per bin. The distribution of the median residual values is sharply peaked, which reflects the fact that normal diploid samples are being analyzed, but has "fat" tails, representing bins with anomalous behavior and those with CNVs (Fig. 2B). We chose to blacklist all bins with

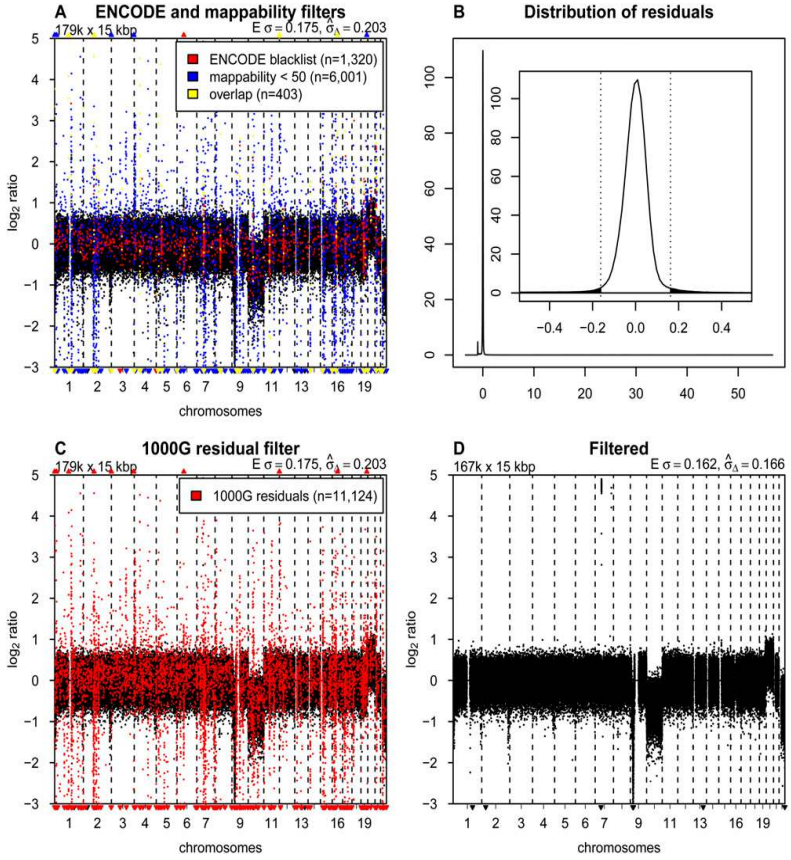

Figure 2. Blacklisting problematic regions. (*A*) Copy number profile for sample LGG150 with bins overlapping with the ENCODE blacklist highlighted in red, bins with mappabilities below 50 highlighted in blue, and the overlap between the two in yellow. (*B*) Distribution of median residuals per bin from the 1000 Genomes Project across the 38 samples. Residuals are defined as the distance between observed read counts and the fitted LOESS surface, divided by the LOESS value. The *outer* plot shows the entire range of values with two discrete peaks. The minor peak around −1.0 results from repetitive sequences. Reads that align equally well to multiple locations in the genome are filtered out. Repetitive sequences therefore have a lower than expected number of reads mapped. The major peak around zero contains most of the bins, and the *inset* shows a magnification of the peak, with the dotted vertical bars and the shaded area showing the cutoff of 4.0 standard deviations (as estimated with a robust first-order estimator) for blacklisting. (*C*) Copy number profile of sample LGG150 with bins in the novel blacklist based on residuals of the 1000 Genomes samples highlighted in red. (*D*) The final copy number profile of sample LGG150 after filtering out bins in the ENCODE and 1000G blacklists.

median residuals greater than 4.0 standard deviations, using a robust first-order estimator (von Neumann et al. 1941) that focuses on the width of the central peak to determine the standard deviation. This procedure removed 10,413 bins. We based our choice of the cutoff on the distributions of residuals found with a number of different bin sizes ranging from 1 to 1000 kb (Supplemental Fig. S5). The cutoff can be adjusted in the QDNAseq package if other values are desired. Changing it by one standard deviation in either direction, however, does not materially affect the results.

We were concerned that the initial presence of bins with high residuals, which were candidates for blacklisting, had the potential to affect the LOESS fit of the initial read counts. Therefore, we implemented an iterative process, recalculating the LOESS correction after removal of the problematic bins found in the previous cycle and again determining the residual distribution. Bins with residuals greater than the same numerical cutoff values established in the first iteration were removed. The list of excluded bins, our blacklist, stabilized at 11,124 bins after 14 iterations. Figure 2C shows the profile of LGG150 with our blacklisted bins highlighted. This blacklist contains many bins not included in the ENCODE list and also includes 97% (6200 of 6404) of the bins with mappabilities below 50. Overlaps between our blacklist, the ENCODE list, and bins with mappabilities below 50 are presented in Supplemental Figure S6. We intentionally were not conservative in blacklisting, since the copy number of a blacklisted locus can be imputed from neighboring bins (assuming no very focal aberrations are present), and most analytical packages handle this imputation automatically (van de Wiel et al. 2011).

For analysis of experimental samples, we routinely remove bins contained in the union of the ENCODE blacklist and our 1000 Genomes-based list at the beginning of the analysis so that their anomalous values do not affect the LOESS GC-mappability fit; although in practice the procedure seems to be fairly robust to the presence of these outliers. Similarly, the LOESS fit could be affected by copy number aberrations present in the data. Therefore, our software allows the correction described in the previous section to be implemented iteratively. After the initial analysis, bins with large LOESS residuals, which presumably are located in copy number aberrations, are excluded and the analysis is repeated. This cycle is iterated until the list of bins that are used stabilizes. We found this approach to be of little benefit in most cases, and the data presented in this paper have been corrected without this iterative step.

In total, this procedure removed 12,278 of the 15-kb bins (6.9%). Together with the 12,893 bins that consist of only uncharacterized nucleotides (N's in the reference genome sequence), they form 954 separate continuous regions, which are listed in Supplemental Table S2. We also list the 2273 genes that fall within these regions, which thus includes genes in common germline CNVs. Figure 2D shows the final profile of sample LGG150 with the blacklist filtering and GC-mappability correction applied. Whole-chromosome losses can be seen involving chromosomes 10 and 22, and a gain of 20. A focal amplification is also present on chromosome 7, as well as a homozygous deletion on 9p. Final profiles for all LGG samples, BT474, AB042, and AB052 are shown in Supplemental Figures S1–S3.

Noise and detection limits

Noise in copy number profiles has contributions from the statistics of counting sequence reads as well as the many steps in the analytical chain from sample acquisition and fixation through DNA isolation, sequencing, and computational processing. Since the variances of independent noise sources are additive, it is convenient to use the variances of the profiles to investigate their noise characteristics. Profiles normalized so that the mean value is 1.0 have variances due to counting statistics equal to $1/N$, where N is the average number of reads per bin (neglecting small effects due to copy number aberrations and the counting corrections). Thus, the difference between the variance of the copy number profile and the variance due to counting statistics ($1/N$) for that profile gives a measure of the noise contribution from the entire sample handling and analytical process, independent of sequence depth. Therefore we examined the dependence of the variances of our profiles on sequence depth.

We first tested the dependence of the variance on read depth alone by subsampling reads from a single data set with 108.4 million mapped reads of sample AB042. The subsampled data ranged over a factor of 100, from about 600 to 6 reads per bin. We performed the subsampling five times at each subsampling level and calculated the variances using a mean-scaled and 0.1%-trimmed first-order estimator. This estimator emphasizes bin-to-bin variation so that it is not affected by copy number aberrations (see Methods). Figure 3A shows the variance of the subsampled data for AB042 versus the variance due to counting statistics ($1/N$). A regression line fitted to the subsampled data has a slope of 1.026 and intercept of 0.00107, very close to the theoretical $1/N$ counting statistics (slope = 1; intercept = 0). The similar behavior of the measured and theoretical slopes indicates that variance versus read depth behaves essentially as expected. The fact that the intercept of the fitted regression line is close to zero indicates that the noise introduced from the sample quality and the analytical process is negligible. Thus, the noise is dominated by counting statistics at the read depths typical for shallow WGS analysis (30 reads per bin). [We note that copy number profiles are typically $\log_2$-transformed in our figures. If the variance were to be calculated based on the transformed profile, the contribution due to read depth would be $\log_2(e)^2/N \approx 2.08/N$].

We also examined the noise introduced by the library preparation and sequencing procedures by performing 10 independent sequencing runs from one DNA isolation of sample AB052. The variances of these profiles are also plotted in Figure 3A. The slope and intercept of the regression line are 1.003 and 0.000781, respectively. Thus, the total variance is again very close to the variance due to counting statistics ($1/N$), indicating that the library preparation and sequencing procedures have an insignificant contribution to the total variance. Further, the profiles from most of the LGG samples and the cell line BT474, which represent completely independent samples, also had variances very close to the theoretical counting statistics limit (Fig. 3A). Thus, DNA samples from a range of specimens obtained from our laboratories provided near optimal data using this measure.

Importantly, the variance characteristics shown in this small set of examples are generally representative of our experience with a large body of clinical specimens from many sources. We have now analyzed over a thousand samples obtained from more than 25 hospitals in five countries, mostly from FFPE tissues. A minority consisted of snap-frozen tissue samples or DNA extracted from cells freshly obtained from peripheral blood, sputum, swabs of the oral mucosa, or cancer cell lines. The samples represented a wide spectrum of neoplasms, mainly carcinomas, but also neuroectodermal and mesenchymal neoplasms, as well as non-neoplastic tissues and cells, generally submitted for detection of somatic aberrations. In most cases, DNAs were isolated in the laboratories that provided the specimens, using their local

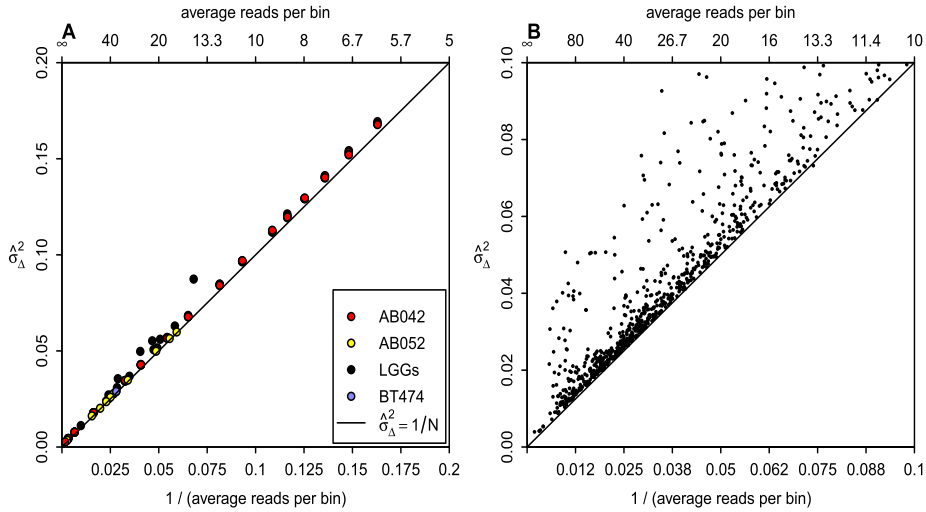

Figure 3. Dependence of variance on sequence depth. (*A*) The relationship between sequence depth and variance $(\hat{\sigma}^2_\Delta)$ for 15 LGGs (black), cell line BT474 (blue), 10 independent library preparations of SCC sample AB052 (yellow), and subsamplings of AB042 data (red). All individual samples are within the *left* half of the graph, with the subsamplings extending to the *right* half as well. The black line shows the linear expectation of the variance as 1/*N*, where *N* is the average number of reads per bin. Lines fitted through the AB042 subsamplings and AB052 repeats have slopes of 1.026 and 1.003, and intercepts of 0.00107 and 0.000781, respectively. (*B*) The relationship between sequence depth and variance for more than a thousand samples sequenced at our institute.

protocols. Samples were sequenced in pools of ~20 per lane. Figure 3B shows the variances of the resulting profiles versus 1/*N* for these samples. This figure shows that for the vast majority of samples, the overall variance in our profiles is dominated by the read depth. In our experience, profiles with a variance corresponding to greater than 30 reads per bin, around 6 million total reads for 15-kb bins (~0.1× sequence coverage), are suitable for most subsequent analyses. Because the noise is dominated by counting statistics for most samples, it is possible to make instructive estimates of the smallest aberrations that can be detected as a function of read depth. In Supplemental Figure S7, we present estimates for gain and loss as well as a simple analytical formula applicable for a wide range of situations.

We note that some samples have variances clearly above the 1/*N* line (Fig. 3B). New sample preparation and analysis of several of these excessively noisy samples indicates that the noise is reproducible, both in magnitude and in shape along the genome, suggesting that it has its origin in the sample. Most likely it is due to degraded/damaged DNA resulting from the fixation and storage. Increasing sequence depth will not reduce this noise relative to the (1/*N*) line for variance due to counting statistics.

The software package QDNAseq

The software package QDNAseq was developed to implement the novel profile correction and blacklisting approach described above and to perform downstream segmentation and calling of aberrations using well established software tools. QDNAseq uses BAM files as input because they are produced by the commonly used alignment algorithms such as BWA (Li and Durbin 2009). The program is implemented in R (R Core Team 2014) and is available in Bioconductor (Gentleman et al. 2004). Detailed information concerning its operation is included in the Bioconductor vignette. Briefly, bin size, LOESS parameters, and blacklisting parameters are adjustable. Blacklisted bins can be visualized, as in Figure 2, A and C. Options are to either filter out bins overlapping with the ENCODE blacklist (1723 bins when using the 15-kb bin size) and/or the blacklist we developed from the 1000G data (11,124 bins). A key feature of QDNAseq is the use of fixed-sized bins, which is necessary for most published downstream procedures that handle series of tumor samples (van de Wiel et al. 2011). Use of fixed-sized bins furthermore allows calculation of annotation data (GC content, mappability, overlap with ENCODE blacklist, 1000G residuals) in advance, facilitating computation and analysis procedures. Analysis is therefore relatively rapid. For example, processing of the LGG150 sample included in this paper takes 75 sec from the input BAM file to the filtered and corrected profile in Figure 2D on a standard workstation or laptop with a 2.3 GHz Intel Core i5 CPU. Included in the QDNAseq package is also an option to compare to matched reference samples should that be desired (see Supplemental Fig. S8).

The output of QDNAseq is read counts per bin, which have been corrected, filtered, normalized, and optionally $\log_2$-transformed. QDNAseq was built in a modular fashion such that analysis tools and pipelines for downstream segmentation and copy number calling previously developed for microarrays (for review, see van de Wiel et al. 2011), for example, can be readily applied. Downstream analysis can also be performed and was tested with the commercially available software suite Nexus Copy Number (BioDiscovery). QDNAseq has also been made available in Chipster (Kallio et al. 2011) and Galaxy (Goecks et al. 2010) and allows export of the copy number results into the Integrative Genomics Viewer (IGV) (Thorvaldsdóttir et al. 2013). The popular segmentation package DNAcopy (Venkatraman and Olshen 2007) can be invoked directly from within QDNAseq. In addition to the existing user-definable parameters available in DNAcopy, an option to smooth signals over a specified number of consecutive bins has been added in QDNAseq. For calling (annotation of segments with copy number states such as gain, amplification, or loss), the package CGHcall (van de Wiel et al. 2007) can be invoked at the user's discretion.

Comparison to other algorithms and array CGH

Multiple algorithms have been developed for DOC DNA copy number analysis. Most compare the tumor sample to a reference signal and thus require acquisition of an appropriate reference sample and additional sequencing. Two algorithms have been published for analysis of shallow WGS that do not require a reference signal, readDepth (Miller et al. 2011) and FREEC (Boeva et al. 2011). Both adjust read counts and/or filter out bins based on GC content and mappability, but lack other blacklisting options such as those based on ENCODE or the 1000 Genomes-based blacklist. Both have integrated segmentation and calling to identify gains and losses. Since the novel aspects of QDNAseq occur in the determination of the filtered and corrected read count profile, we opted to evaluate the performance of QDNAseq relative to the preprocessing parts of these other analysis packages. However, readDepth does not output bin-level data so we could only compare our results with FREEC. A third program, CLImAT, was recently published which, among other things, infers copy number from the observed sequence depth without requiring a reference signal (Yu et al. 2014). The goal of this program, however, is to use relatively deep (10× genome coverage) sequencing to obtain information that is not available from a small number of reads (0.1× genome coverage), which is the focus of our work. The CLImAT algorithm uses a simpler form of simultaneous GC and mappability correction that is likely to be too noisy at our read depth, so we did not evaluate it.

Both QDNAseq and FREEC perform better than the Agilent array CGH platform at the sequencing depths used here. Figure 4, A, B, and C shows the profiles of sample LGG150 obtained with QDNAseq, array CGH, and FREEC, respectively. The data from QDNAseq and FREEC are very similar in their calculated noise, but FREEC contains several focal apparent gains and losses that are not present in the QDNAseq data due to blacklist filtering. These artifactual features in FREEC output are at risk of being interpreted as true aberrations. Array CGH has greater noise and more outliers than with both sequencing analyses, with a standard deviation of 0.19 compared to 0.17 for sequencing analyses. Moreover, the deflections for the copy number changes are larger for the sequencing methods than for array CGH. The average signal-to-noise ratio (SNR) for 12 whole-chromosome aberrations among

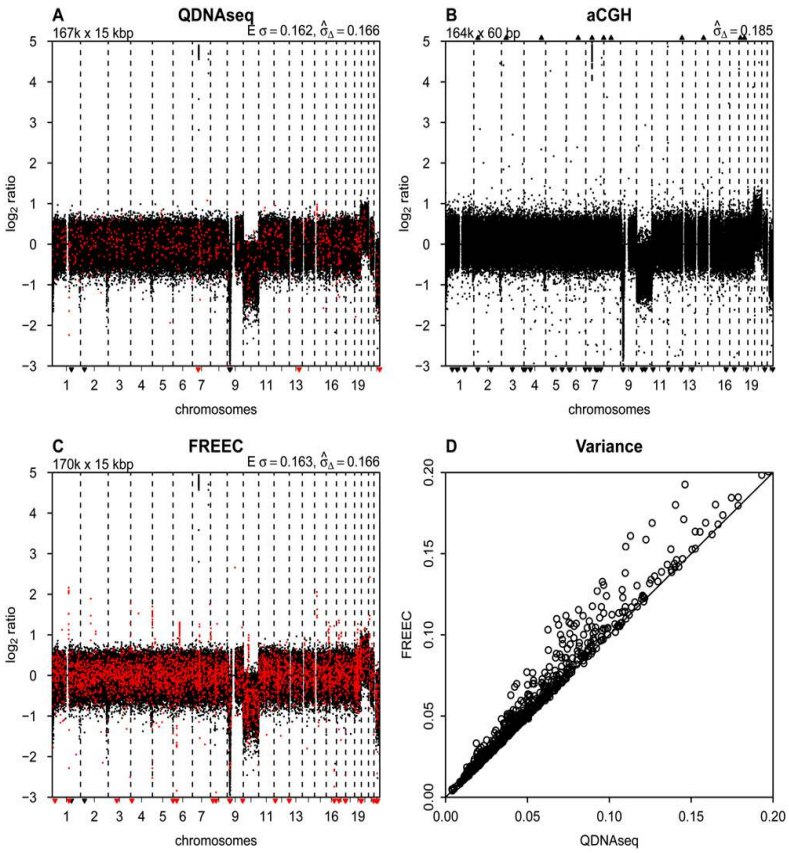

Figure 4. Comparison to other methods. (*A*) Final copy number profile of sample LGG150 obtained with QDNAseq after removing blacklisted bins and correcting read counts for GC content and mappability. This procedure results in 166,909 bins, and highlighted in red are those 750 bins that are not contained in the output of FREEC. (*B*) Copy number profile of sample LGG150 obtained with an Agilent 180K microarray with 164,378 unique array elements. (*C*) Copy number profile of sample LGG150 obtained with FREEC with 170,474 bins. Highlighted in red are those 4315 bins that are not contained in the output of QDNAseq. Note that many of the red bins are in focal peaks that have the potential of being called aberrations but which are probably spurious since they are contained in the QDNAseq blacklists. (*D*) Noise ($\hat{\sigma}^2_\Delta$) for QDNAseq versus FREEC calculated from the thousand samples in Figure 3B. Only the 166,159 bins present in the output of both algorithms were used in order to eliminate differences caused by blacklisting spurious bins.

the 15 LGG samples was 1.89, 1.91, and 1.40 for QDNAseq, FREEC, and aCGH, respectively (Supplemental Table S3).

Comparison of data obtained with QDNAseq and FREEC on the entire set of ~1000 samples shows that the variance with FREEC was never lower than with QDNAseq and often somewhat higher (Fig. 4D). To assure that the comparison concentrates on differences in read count corrections and not the filtering of blacklisted bins, the variance calculations were performed only on the set of bins contained in the output of both programs. This shows that the simultaneous correction for GC content and mappability implemented in QDNAseq always performs at least as well as the sequential corrections in FREEC and is better for some samples.

Simultaneous correction for GC content and mappability outperforms separate corrections for cases in which the two parameters interact. The interaction can be seen from examination of the LOESS surfaces for the various samples. For sample LGG150 presented in Figure 1, read counts always increase with increasing mappability regardless of GC content. Thus to a reasonable approximation there is minimal interaction. In contrast, samples LGG155 and LGG259 both have read count maxima along the mappability-axis that vary with GC content (Supplemental Figs. S1, S2). Consequently, a single correction curve for GC that is applied to all mappabilities, and vice versa, will not properly correct these samples.

The major benefit of our simultaneous correction approach is seen in the removal of spurious regions of variation in the profiles. Supplemental Figure S9 shows profiles for the whole genome and for chromosome 1 generated from the same sequencing data by QDNAseq and FREEC for all 15 LGG samples, two SCCs, and BT474. Examination of the profiles for LGG151 and LGG155 shows clearly that small features in the profiles produced by FREEC are corrected by QDNAseq. Similar features also occur on other chromosomes. Improved correction of this sample-related variability allows use of more sensitive segmentation and calling procedures for a given level of false positives. Three of the 15 LGG samples showed significant improvement using QDNAseq. Thus our correction procedure facilitates correct biological interpretations from samples with a wider range of quality.

Discussion

We have described a shallow WGS procedure designed to obtain high-quality DNA copy number information from fresh and archival samples. The method was developed in the process of analyzing over a thousand tumor DNA samples obtained from more than 25 hospitals in five countries, mostly from FFPE tissue. Our goal was to provide the best possible read count profiles so that subsequent segmentation and calling steps would be able to sensitively detect true aberrations at acceptable levels of false positives. The data presented show that our corrected profiles have noise levels very near the fundamental limit imposed by the statistics of read counting for most samples, and are less sensitive to DNA quality induced artifacts than profiles produced by prior approaches. The predictable nature of the major noise source of our read count profiles represent a considerable interpretive simplification compared to microarray DNA copy number profiles, in which the noise sources are obscure and copy number changes are frequently reduced in magnitude due to array performance (Snijders et al. 2001; Ylstra et al. 2006).

Our procedure contains two novel features: simultaneous correction of counts for GC content and mappability, and empirical recognition of problematic regions of the genome based on analysis of a group of normal samples. Here, we demonstrate the performance of QDNAseq on 1000 samples, mostly from archival FFPE cases. The simultaneous correction for GC and mappability, using a LOESS fit of the raw count data to the average values of these parameters for each sequencing bin, always performs at least as well, and in more degraded DNA samples better, than the separate corrections that are used by most existing algorithms. Nevertheless, it is also evident that our correction remains inadequate for some samples. It is likely that a more thorough understanding of the impact of formalin fixation and the distribution of base composition and mappabilities within the sequencing bins will result in improved ability to obtain useful copy number data from samples that remain problematic. Further, although our blacklist was developed from 38 normal samples representing a variety of ethnicities from the 1000 Genomes Project, similarly derived blacklists tailored to the ethnicity of the population from which the samples were obtained would allow more precise blacklisting of common germ-line CNVs relevant for that population.

Shallow WGS is cost effective. Our experience indicates that high quality DNA copy number information can be obtained with ~6 million reads per sample (~0.1× genome coverage). High-capacity instruments such as the Illumina HiSeq can obtain this read depth with a multiplex analysis of 20 samples per lane. We achieved a further increase in efficiency by sequencing only 50 bp from one end of the DNA molecules, which also allows the use of compromised samples with short DNA fragments. This level of sequence depth provides better resolution than is available from microarrays and costs significantly less. Shallow sequencing has a particular cost benefit if combined with exome sequencing because the initial preparation for both is the same. The additional cost of the shallow sequence run is marginal (~5% extra) and provides high-resolution genome-wide copy number information. If the shallow sequencing run is performed prior to exome enrichment and the exome sequencing run, it can also serve as the ultimate quality control. Although we obtained most of our data on an Illumina HiSeq instrument, the use of smaller capacity sequencers, such as the Illumina MiSeq, offer rapid turnaround and have the required capacity for relatively infrequently submitted diagnostic samples.

Methods

Sample selection

Fifteen LGGs (van Thuijl et al. 2014), two SCCs (Bhattacharya et al. 2011), and the breast cancer cell line BT474 were used to develop and illustrate the shallow WGS pipeline presented. All material used from LGG and SCC tumors was derived from FFPE archival samples. Patient consent was obtained for SCCs as published previously (Bhattacharya et al. 2011). LGG samples were collected from five Dutch hospitals (VU University Medical Center in Amsterdam, Academic Medical Center in Amsterdam, Radboud University Medical Center in Nijmegen, St. Elisabeth Hospital in Tilburg, and Isala Klinieken in Zwolle). Sample collection was approved by the Medical Ethics Committees of all five hospitals. Areas containing > 60% tumor cells were outlined on hematoxylin and eosin-stained slides, and 10 subsequent 10-μm sections were used for DNA isolations.

DNA from the LGG samples was isolated as previously described (van Essen and Ylstra 2012). DNA concentrations were measured with the Nanodrop 2000 (Fisher Scientific), and 500 ng was used as input for Shallow WGS laboratory preparation. DNA from the SCC samples was isolated as previously described (Bhattacharya et al. 2011),

DNA concentrations were measured with the Qubit 2.0 fluorometer dsDNA BR Assay (Life Technologies), and 250 ng DNA used as input for shallow WGS laboratory preparation. The BT474 breast tumor cell line was cultured and DNA isolated as previously described (Krijgsman et al. 2013). DNA concentration was measured with the Qubit fluorometer and 250 ng used as input.

Shallow WGS laboratory preparation

DNA was sheared on a Covaris S2 (Covaris) with the following settings: duty cycle 10%, intensity 5.0, bursts per sec 200, duration 240 sec (FFPE), duration 300 sec (fresh and fresh-frozen), mode frequency sweeping, power 23V, temperature 5.5°C to 6°C, water level 15. Sample preparation was then performed with the TruSeq DNA kit V2 (Illumina). After end repair and 3′ adenylation, adapter ligation was performed with 1 μL of adapter index for fresh (frozen) samples and 0.55 μL of adapter index for FFPE samples. Final sequence library amplification was performed with 10 PCR cycles for FFPE derived DNA samples or eight cycles for DNA derived from fresh or fresh-frozen samples. One PCR cycle included 10 sec 98°C, 30 sec 60°C, and 30 sec 72°C. The PCR program started with 30 sec 98°C and ended with 5 min 72°C. The final holding temperature was 10°C.

The yield of the sequence library was assessed with a Bioanalyzer DNA 1000 and/or HS DNA (Agilent Technologies). Libraries with small PCR products (~120 nt in length caused by unligated adapter dimers) or large PCR products (> 1000 nt in length caused by an exhausted PCR mix) were selected for cleaning. Cleaning was performed by using a double-sided bead size selection procedure with Agencourt AMPure XP beads (Beckman Coultier). Libraries were equimolarly pooled with 18–22 barcoded samples and 7 pM molarity loaded per lane of a HiSeq Single End Flowcell (Illumina). This was followed by cluster generation on a cBot (Illumina) and sequencing on a HiSeq 2000 (Illumina) in a single-read 50-cycle run mode (SR50).

Alignment to reference genome

Sequence reads were aligned to the human reference genome build GRCh37/hg19 downloaded from Ensembl (Flicek et al. 2013) with BWA 0.5.9 (Li and Durbin 2009), with a maximum edit distance of 2 and base trimming quality of 40. PCR duplicates were marked with Picard 1.61 (http://broadinstitute.github.io/picard/), and filtered out with SAMtools 0.1.18 (Li et al. 2009) together with reads with mapping qualities (MAPQ) lower than 37. We note that the maximum possible value and the distribution of mapping qualities varies between aligners, and a different cutoff might be suitable for e.g., Bowtie (Langmead and Salzberg 2012), which was tested here but did not show an improvement over BWA for copy number assessment.

Annotations for genomic bins

The genome was divided into nonoverlapping, fixed-sized bins of 15 kb. GC content of each bin was calculated as number of C and G nucleotides divided by number of A, C, G, and T nucleotides in the reference sequence. The percentage of characterized nucleotides was calculated by dividing the number of nucleotides A, C, G, and T with the bin size (15 kb). This is used to adjust read counts for bins partially covered by uncharacterized nucleotides (N's) or incomplete bins at the very ends of chromosomes.

Mappability is a measure of the uniqueness of a specific sequence in the reference genome and depends on the length of the sequence and the number of mismatches allowed. If $F_k(x)$ is the frequency at which the k-mer sequence at position x is observed in the reference genome sequence and its reverse complement, the mappability of this position is defined as $M_k(x) = 1/F_k(x)$. In this paper, we use the term mappability to refer to the average mappability of all 50-mer sequences within a bin, allowing for two mismatches, and scaling the value from 0 to 100. These values were calculated from the ENCODE alignability track for 50-mers (data version January 2010) with the *bigWigAverageOverBed* program downloaded from the UCSC Genome Browser (Kent et al. 2002; Rosenbloom et al. 2013).

The ENCODE blacklisted regions (March 2012 Freeze) were used to calculate percent overlap with each bin. Pregenerated bin annotations are available for human reference genome build GRCh37/hg19 and bin sizes of 1, 5, 10, 15, 30, 50, 100, 500, and 1000 kb.

Binning and correction of read counts

The number of sequence reads in each bin was calculated. Read counts were adjusted for those 487 bins that are only partly covered by characterized nucleotides in the reference genome sequence. For a bin containing a proportion r of uncharacterized nucleotides, no reads could be mapped to this fraction of the bin, and the read count was therefore adjusted by dividing it by $1 - r$.

Next, median read counts were calculated as a function of GC content and mappability. For this purpose, GC content and mappability values were rounded to integers (IEC 60559 standard). A two-dimensional LOESS surface was then fitted to the observed median read counts. The read count for each bin was then corrected by dividing it with the fitted LOESS value.

Optimal parameters for the LOESS correction were evaluated with an odd-even cross-validation as follows: Bins were divided into odd and even bins, and only odd ones were used to calculate the LOESS correction as above. The same correction was then applied to both odd and even bins, and the absolute values for differences in adjusted counts between adjacent odd and even bins were calculated. A test statistic was calculated as a trimmed mean of the absolute values after removal of the upper 10% to account for copy number breakpoints. Parameter values of span = 0.65 and family = 'symmetric' were chosen to minimize the value of the test statistic.

1000 Genomes residuals

The blacklist based on 1000 Genomes samples was generated as follows. Publicly available samples from the 1000 Genomes Project that matched the experimental setup were downloaded (Illumina single-read of at least 50 bp, low coverage, whole-genome sequencing). For samples with read lengths longer than 50 bp, the reads were truncated to the first 50 bp. In total, 38 samples that matched the experimental setup were available. Alignment and two-dimensional LOESS correction were then performed as outlined above. Residuals [(observed read count − LOESS fit)/LOESS fit] from the correction were recorded, and medians per bin were then calculated across the 38 samples. Cutoff for exclusion was set at 4.0 standard deviations (as estimated with a robust first-order estimator [von Neumann et al. 1941]). After bins exceeding this cutoff were removed, the LOESS correction was repeated without these anomalous bins and residuals calculated again. This process was repeated until the list of bins to be excluded stabilized.

Noise, comparison to other algorithms, and array CGH

FREEC (version 6.4) (Boeva et al. 2011) was run with a bin size of 15 kb, mappability-based read count correction turned on, minimum mappability set to 50, and otherwise default settings. These

settings were selected to mimic QDNAseq as closely as possible. As a measure of noise, we used an estimator based on first-order differences (von Neumann et al. 1941). This noise estimator is sensitive to uncorrelated bin-to-bin read count differences along the profile, but is largely unaffected by correlated behavior of groups of bins such as steps in the profile due to true copy number aberrations or long-range waviness. For robustness against large outliers, we excluded 0.1% of extreme values from both ends of the distribution. Unless specified otherwise, terms "standard deviation" and "variance" used in this paper refer to this mean-scaled 0.1%-trimmed first-order estimate and its square, respectively. They are calculated for a linear representation of the profiles even though we present $\log_2$-transformed profiles for display convenience. The standard deviation of a profile, denoted by $\hat{\sigma}_\Delta$, and the theoretically expected standard deviation based on read counting, denoted by E σ, are given above each profile.

All samples were profiled with CGH arrays that contained 180K in situ synthesized 60-mer oligonucleotides evenly distributed (every 17 kb) across the genome (Agilent Technologies). BT474 CGH arrays were performed previously (Krijgsman et al. 2013) and data downloaded from the GEO database (Edgar et al. 2002) with accession number GSM903069. Labeling, hybridization, scanning, and feature extraction were carried out as previously described (Krijgsman et al. 2013) with pooled normal reference samples. After median normalization, wave-correction was performed with NoWaves (van de Wiel et al. 2009), which would account for GC variation across the genome. The SCC samples have also been previously characterized by 2K BAC arrays which data are available in GEO with accession GSE28407 (Bhattacharya et al. 2011).

Data access

QDNAseq package is available through Bioconductor (http://www.bioconductor.org/) (Gentleman et al. 2004). Source code is available in GitHub (https://github.com/ccagc/QDNAseq/), and for the version used to generate data presented in this paper, also in the Supplemental Material. Sequence and microarray data have been submitted to the European Genome-phenome Archive (EGA; http://www.ebi.ac.uk/ega/) (Leinonen et al. 2011) which is hosted at the EBI, under accession number EGAS00001000642.

Acknowledgments

We thank colleagues at the Leeds Institute of Molecular Medicine (UK) for fruitful discussions. We thank Francisco Real (CNIO), James Brenton (University of Cambridge), Jan Molenaar (AMC), and VUmc colleagues for their permission to use "noise vs. sequence depth" of the samples submitted by their institute in Figure 3B. This work was supported by the Dutch Cancer Society (KWF) grant 2009-4470; Stichting STOPhersentumoren.nl; VUmc Cancer Center Amsterdam (VUmc CCA); the Center for Translational Molecular Medicine (CTMM) projects 03O-101 DeCoDe and 05T-401 ICT TraIT; the Centre for Medical Systems Biology (CMSB-NWO); and the European Community's Framework Programme Seven (FP7) under contract no. 278981 "AngioPredict." Funding from the National Institutes of Health (NIH grants CA089715, CA113833, CA131286, and CA163019) also supported the work. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. Part of this work was carried out on the Dutch national e-infrastructure with the support of the SURF Foundation. The 1000 Genomes and ENCODE Projects are acknowledged for the data they have generated. The ENCODE mappability data were generated in Roderic Guigó's laboratory at the Centre for Genomic Regulation (CRG), Barcelona, Spain, and the blacklist data at Duke University's Institute for Genome Sciences and Policy (IGSP), University of Cambridge, Department of Oncology and CR-UK Cambridge Research Institute (CRI), and Stanford University in cooperation with the European Bioinformatics Institute (EBI).

Author contributions: I.S. and D.S. developed laboratory procedures and algorithms, implemented the QDNAseq software package, and wrote the manuscript; H.B. provided the statistical framework and implemented the QDNAseq software package; M.A.v.d.W. and A.B.O. provided the statistical framework; H.F.v.E., P.P.E., and F.R. performed array hybridizations and WGS DNA preparations; H.F.v.T., G.A.M., J.C.R., and P.W. contributed samples and pathology revisions; D.P. and D.G.A. contributed samples and wrote the manuscript; B.Y. conceived the project and wrote the manuscript.

References

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65.

Baldwin EL, Lee JY, Blake DM, Bunke BP, Alexander CR, Kogan AL, Ledbetter DH, Martin CL. 2008. Enhanced detection of clinically relevant genomic imbalances using a targeted plus whole genome oligonucleotide microarray. *Genet Med* **10:** 415–429.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* **40:** e72.

Bhattacharya A, Roy R, Snijders AM, Hamilton G, Paquette J, Tokuyasu T, Bengtsson H, Jordan RCK, Olshen AB, Pinkel D, et al. 2011. Two distinct routes to oral cancer differing in genome instability and risk for cervical node metastasis. *Clin Cancer Res* **17:** 7024–7034.

Blow N. 2007. Tissue preparation: tissue issues. *Nature* **448:** 959–963.

Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, Barillot E. 2011. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics* **27:** 268–269.

Casparie M, Tiebosch ATMG, Burger G, Blauwgeers H, van de Pol A, van Krieken JHJM, Meijer GA. 2007. Pathology databanking and biobanking in The Netherlands, a central role for PALGA, the nationwide histopathology and cytopathology data network and archive. *Cell Oncol* **29:** 19–24.

Chiang DY, Getz G, Jaffe DB, O'Kelly MJT, Zhao X, Carter SL, Russ C, Nusbaum C, Meyerson M, Lander ES, et al. 2009. High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* **6:** 99–103.

Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast computation and applications of genome mappability. *PLoS ONE* **7:** e30377.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* **30:** 207–210.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Feuk L, Carson AR, Scherer SW. 2006. Structural variation in the human genome. *Nat Rev Genet* **7:** 85–97.

Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. 2013. Ensembl 2013. *Nucleic Acids Res* **41:** D48–D55.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5:** R80.

Goecks J, Nekrutenko A, Taylor J, Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* **11:** R86.

Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S. 2012. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* **28:** 40–47.

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. *Cell* **144:** 646–674.

Hostetter G, Kim SY, Savage S, Gooden GC, Barrett M, Zhang J, Alla L, Watanabe A, Einspahr J, Prasad A, et al. 2010. Random DNA fragmentation allows detection of single-copy, single-exon alterations of copy number by oligonucleotide array CGH in clinical FFPE samples. *Nucleic Acids Res* **38:** e9.

Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, Koski M, Käki J, Korpelainen EI. 2011. Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics* **12:** 507.

Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. 1992. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258:** 818–821.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996–1006.

Krijgsman O, Israeli D, Haan JC, van Essen HF, Smeets SJ, Eijk PP, Steenbergen RDM, Kok K, Tejpar S, Meijer GA, et al. 2012. CGH arrays compared for DNA isolated from formalin-fixed, paraffin-embedded material. *Genes Chromosomes Cancer* **51:** 344–352.

Krijgsman O, Israeli D, van Essen HF, Eijk PP, Berens MLM, Mellink CHM, Nieuwint AW, Weiss MM, Steenbergen RDM, Meijer GA, et al. 2013. Detection limits of DNA copy number alterations in heterogeneous cell populations. *Cell Oncol* **36:** 27–36.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359.

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, et al. 2011. The European Nucleotide Archive. *Nucleic Acids Res* **39:** D28–D31.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25:** 1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* **20:** 265–272.

McSherry EA, Mc Goldrick A, Kay EW, Hopkins AM, Gallagher WM, Dervan PA. 2007. Formalin-fixed paraffin-embedded clinical tissues show spurious copy number changes in array-CGH profiles. *Clin Genet* **72:** 441–447.

Miller CA, Hampton O, Coarfa C, Milosavljevic A. 2011. ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS ONE* **6:** e16327.

Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA, et al. 2008. The fine-scale and complex architecture of human copy-number variation. *Am J Hum Genet* **82:** 685–695.

Pinkel D, Segraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, et al. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet* **20:** 207–211.

R Core Team. 2014. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, Hutter B, Jäger N, Kool M, Taylor M, Lichter P, et al. 2013. Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PLoS ONE* **8:** e66621.

Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, Wong MC, Maddren M, Fang R, Heitner SG, et al. 2013. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res* **41:** D56–D63.

Shen JJ, Zhang NR. 2012. Change-point model on nonhomogeneous Poisson processes with application in copy number profiling by next-generation DNA sequencing. *Ann Appl Stat* **6:** 476–496.

Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, et al. 2001. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet* **29:** 263–264.

Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. 2012. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* **28:** 2711–2718.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14:** 178–192.

van Thuijl HF, Ylstra B, Würdinger T, van Nieuwenhuizen D, Heimans JJ, Wesseling P, Reijneveld JC. 2012. Genetics and pharmacogenomics of diffuse gliomas. *Pharmacol Ther* **137:** 78–88.

van Thuijl HF, Scheinin I, Sie D, Alentorn A, van Essen HF, Cordes M, Fleischeuer R, Gijtenbeek AM, Beute G, van den Brink WA, et al. 2014. Spatial and temporal evolution of distal 10q deletion; a prognostically unfavorable event in diffuse low-grade gliomas. *Genome Biol* **15:** 471.

van de Wiel MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, Ylstra B. 2007. CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics* **23:** 892–894.

van de Wiel MA, Brosens R, Eilers PHC, Kumps C, Meijer GA, Menten B, Sistermans E, Speleman F, Timmerman ME, Ylstra B, et al. 2009. Smoothing waves in array CGH tumor profiles. *Bioinformatics* **25:** 1099–1104.

van de Wiel MA, Picard F, van Wieringen WN, Ylstra B. 2011. Preprocessing and downstream analysis of microarray DNA copy number profiles. *Brief Bioinform* **12:** 10–21.

van Essen HF, Ylstra B. 2012. High-resolution copy number profiling by array CGH using DNA isolated from formalin-fixed, paraffin-embedded tissues. *Methods Mol Biol* **838:** 329–341.

Venkatraman ES, Olshen AB. 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics* **23:** 657–663.

von Neumann J, Kent R, Bellinson H, Hart B. 1941. The mean square successive difference. *Ann Math Stat* **12:** 153–162.

Warren KE, Killian K, Suuriniemi M, Wang Y, Quezado M, Meltzer PS. 2012. Genomic aberrations in pediatric diffuse intrinsic pontine gliomas. *Neuro Oncol* **14:** 326–332.

Xi R, Hadjipanayis AG, Luquette LJ, Kim TM, Lee E, Zhang J, Johnson MD, Muzny DM, Wheeler DA, Gibbs RA, et al. 2011. Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci* **108:** E1128–E1136.

Xie C, Tammi MT. 2009. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10:** 80.

Ylstra B, van den Ijssel P, Carvalho B, Brakenhoff RH, Meijer GA. 2006. BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res* **34:** 445–450.

Yu Z, Liu Y, Shen Y, Wang M, Li A. 2014. CLImAT: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data. *Bioinformatics* **30:** 2576–2583.

Received March 17, 2014; accepted in revised form September 15, 2014.

Chapter 5

Spatial and temporal evolution of distal 10q deletion, a prognostically unfavorable event in diffuse low-grade gliomas

Hinke F van Thuijl[1]
Ilari Scheinin[1]
Daoud Sie
Agusti Alentorn
Hendrik F van Essen
Martijn Cordes
Ruth Fleischeuer
Anja M Gijtenbeek
Guus Beute
Wimar A van den Brink
Gerrit A Meijer
Miek Havenith
Ahmed Idbaih
Khê Hoang-Xuan
Karima Mokhtari
Roel GW Verhaak
Paul van der Valk
Mark A van de Wiel
Jan J Heimans
Eleonora Aronica
Jaap C Reijneveld
Pieter Wesseling
Bauke Ylstra

[1] Shared first authors.

This publication has also been included in Hinke van Thuijl's dissertation, **Molecular characterization of low-grade glial neoplasms**, VU University Amsterdam, 2016.

Genome Biology (2014) **15:** 471–483

Genome Biology

RESEARCH Open Access

Spatial and temporal evolution of distal 10q deletion, a prognostically unfavorable event in diffuse low-grade gliomas

Hinke F van Thuijl[1,2†], Ilari Scheinin[1,3†], Daoud Sie[1], Agusti Alentorn[4,5,6,7], Hendrik F van Essen[1], Martijn Cordes[1], Ruth Fleischeuer[8], Anja M Gijtenbeek[9], Guus Beute[10], Wimar A van den Brink[11], Gerrit A Meijer[1], Miek Havenith[12], Ahmed Idbaih[4,5,6,7], Khê Hoang-Xuan[4,5,6,7], Karima Mokhtari[4,5,6,13], Roel GW Verhaak[14,15], Paul van der Valk[1], Mark A van de Wiel[16,17], Jan J Heimans[2], Eleonora Aronica[18], Jaap C Reijneveld[2,19], Pieter Wesseling[1,20] and Bauke Ylstra[1*]

Abstract

Background: The disease course of patients with diffuse low-grade glioma is notoriously unpredictable. Temporal and spatially distinct samples may provide insight into the evolution of clinically relevant copy number aberrations (CNAs). The purpose of this study is to identify CNAs that are indicative of aggressive tumor behavior and can thereby complement the prognostically favorable 1p/19q co-deletion.

Results: Genome-wide, 50 base pair single-end sequencing was performed to detect CNAs in a clinically well-characterized cohort of 98 formalin-fixed paraffin-embedded low-grade gliomas. CNAs are correlated with overall survival as an endpoint. Seventy-five additional samples from spatially distinct regions and paired recurrent tumors of the discovery cohort were analyzed to interrogate the intratumoral heterogeneity and spatial evolution. Loss of 10q25.2-qter is a frequent subclonal event and significantly correlates with an unfavorable prognosis. A significant correlation is furthermore observed in a validation set of 126 and confirmation set of 184 patients. Loss of 10q25.2-qter arises in a longitudinal manner in paired recurrent tumor specimens, whereas the prognostically favorable 1p/19q co-deletion is the only CNA that is stable across spatial regions and recurrent tumors.

Conclusions: CNAs in low-grade gliomas display extensive intratumoral heterogeneity. Distal loss of 10q is a late onset event and a marker for reduced overall survival in low-grade glioma patients. Intratumoral heterogeneity and higher frequencies of distal 10q loss in recurrences suggest this event is involved in outgrowth to the recurrent tumor.

Background

Diffuse low-grade gliomas (LGGs) are regarded as slow growing malignant brain tumors and patients can live up to 30 years with this disease. In a subset of patients the tumor exerts a more aggressive behavior and survival can be as short as two years [1]. Personalized timing of postoperative treatment is crucial to forestall progression in the latter group whilst preventing long-term side-effects for patients with more favorable prospects [2]. The disease course of patients with LGGs is correlated with gene mutations, such as in *p53* and *IDH1*, hypermethylation of *MGMT* as well as chromosomal copy number aberrations (CNAs). Regarding the latter, assessment of combined loss of 1p and 19q currently is implemented in routine clinical care in specific glioma subgroups given its favorable prognostic and predictive value [3,4]. Other CNAs, such as losses of chromosomes 10 and 11p, have been reported to be prognostically unfavorable, but have not been introduced into clinical practice yet, possibly due the limited number of samples included in the studies and/or lack of validation [5-7]. Unfavorable events might go undetected as a consequence of intratumoral

* Correspondence: b.ylstra@vumc.nl
†Equal contributors
[1]Department of Pathology, VU University Medical Center, 1007 MB Amsterdam, The Netherlands
Full list of author information is available at the end of the article

heterogeneity in gliomas [8], which is particularly salient if they are only present in the more malignant subclones of LGGs that seed outgrowth of a recurrent tumor [9], thereby promoting a large extent of resection. As current knowledge on the temporal and spatial evolution of CNAs in LGGs is limited, we evaluated CNAs in a clinically and histologically representative cohort of formalin-fixed archival samples using shallow whole genome sequencing (shallow WGS). We demonstrate that loss of part or whole chromosome 10q is prognostically unfavorable and often present in a subclonal manner.

Results

Clinical and histological data

We studied 173 formalin-fixed paraffin-embedded (FFPE) samples from 98 LGG patients, including spatially distinct regions and paired recurrent tumors, by shallow WGS. Patients had either deceased or had passed the median survival time of six years for LGGs. Other inclusion criteria and patient characteristics of this discovery cohort are summarized in Figure 1 and Table 1. Age at diagnosis, overall survival and postoperative treatment (type and timing) varied extensively between patients, but not between the five participating hospitals, which contributed nearly equal numbers of cases. Comparison of overall survival between patients treated immediately after surgery and those for whom postoperative treatment was withheld did not reveal statistically significant differences. Characteristics of the LGG patients of the French validation (n =126) [5] and confirmation cohorts (n =184), the latter from publicly available data from The Cancer Genome Atlas (TCGA) [10], are also listed in Table 1. Due to the retrospective character of this study, the cohorts are not matched; there is considerable variation in duration of follow-up and the percentage of patients deceased and for which information on overall survival is available, as well as in the distribution of histological subgroups.

Copy number detection by shallow WGS in LGGs

To obtain genome-wide copy numbers from the FFPE samples of our discovery cohort, we evaluated the use of shallow WGS. First, for sample LGG284 a paired-end 100 (PE100) sequence run was performed. Copy number profiles were produced by counting the unique sequence

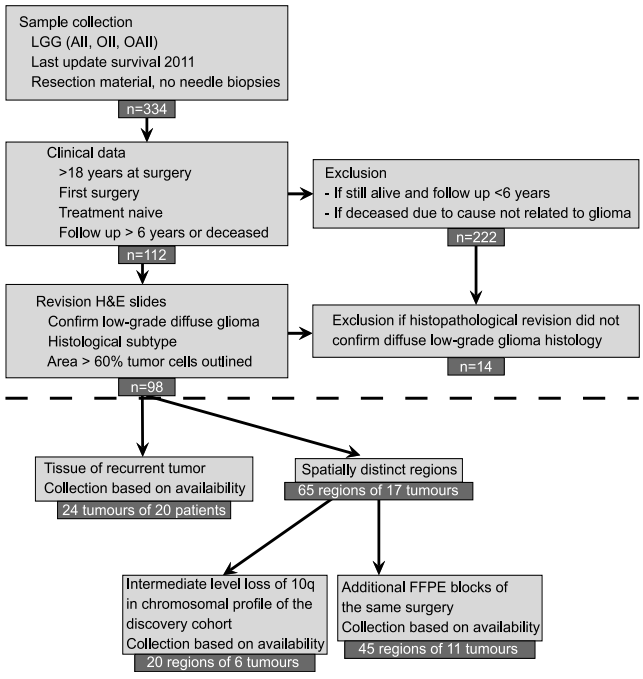

Figure 1 Accrual of samples and clinical data of the discovery cohort. Samples were selected based on criteria listed in the boxes connected with vertical arrows. Reasons for exclusion of samples are listed in boxes in the right panel, connected with horizontal arrows. The number below each box represents the number of patients. Boxes below the dotted line list criteria for collection of recurrent tumors and spatially distinct regions of LGGs. AII, astrocytoma; H&E, hematoxylin and eosin; OII, oligodendroglioma; OAII, oligoastrocytoma.

Table 1 Characteristics of diffuse low-grade glioma patients in the discovery, validation and confirmation cohorts

Variable	Dutch discovery cohort (n =98)	French validation cohort (n =126)	TCGA confirmation cohort (n =184)
Gender			
Female	50 (51%)	59 (47%)	93 (51%)
Age at diagnosis (years)			
Mean	40.3	39.8	40.8
Median	39.6	39.4	39
Range	(21-83)	(18-76)	(14-87)
Duration of follow-up of patients still alive at last evaluation (months)			
Mean	133.6	39.6	23.7
Median	129.0	30.9	10.7
Range	72-288	1-187	1-185
Patients deceased	46 (47%)	32 (25%)	18 (10%)
Overall survival of patients deceased at last evaluation (months)			
Mean	89	48.6	63.5
Median	149.0	51.8	65.6
Range	1-361	0.1-98	1.2-132.6
Histological subtypes			
Oligodendroglioma	43 (43%)	53 (42%)	87 (47%)
Astrocytoma	42 (42%)	23 (18%)	40 (22%)
Oligoastrocytoma	15 (15%)	50 (40%)	57 (31%)

tags per 15 kb bin of the paired-end 100 bp reads from both ends (PE100 in Figure S1A in Additional file 1), the single 100 bp read from one end (SR100 in Figure S1B in Additional file 1) and the trimmed first single 50 bp read from the same end (SR50 in Figure S1C in Additional file 1). The noise (measured as variance) of the different profiles is very similar and CNAs observed are indistinguishable from each other, which implies that the uniqueness of the 50 bp sequence tags suffices to infer copy number levels, and longer reads are not necessary. Array comparative genomic hybridization (array CGH) was performed on the same DNA sample, which confirmed the CNAs detected (Figure S1D in Additional file 1). For an additional eight samples both 50 bp single-read (SR50) shallow WGS and array CGH were applied as technical validation. Shallow WGS and array analysis invariably yielded the same CNA profiles (Figure S2 in Additional file 1). Based on this information, all subsequent analyses were performed using 50 bp single-read (SR50) shallow WGS since it is more cost-effective and allows the use of samples with short DNA fragments, which are frequently obtained with FFPE materials. The most frequent CNAs, detected in more than 10% of cases, are whole or partial loss of chromosomal arms 9p, 10q, 12p, 13 and 14, as well as gain of chromosomal arms 7q, 8q, 10p and 11q. The most frequent CNAs in this cohort are co-deletion of 1p and 19q often accompanied by loss of whole chromosome 4, all commensurate with previous reports [11] (Figure 2).

The prognostic value of CNAs in discovery, validation and confirmation cohorts

Association of survival with CNAs detected in the discovery cohort was tested. In addition to the known prognostically favorable 1p/19q co-deletion, five further chromosomal losses at chromosomes 9p, distal 10q, 11p, 13q, and 22q presented with statistical significance (Table 2). No associations were observed with gains. Significant regions were verified in the French validation cohort of 126 diffuse LGG patients (Table 1). Loss of distal 10q was an unfavorable CNA in both cohorts, whereas losses of chromosomal regions at 9p, 11p, 13q and 22q were not substantiated in the validation cohort (Table 2). In the discovery cohort, median overall survival for patients with or without loss of whole or distal 10q was respectively 6.6 years versus 16.7 years (18/98, *P*-value =0.009). The size of chromosome 10 deletion varies from whole chromosome loss (5/18) to 22.5 Mbp distal loss (10q25.2- 10qter). An association between loss of this region with overall survival was finally tested in the TCGA dataset of LGG. Despite the limited number of patients deceased in this cohort (Table 1), a significant association with overall survival was observed (*P*-value =0.0018) (Figure 3), which confirms that distal 10q is a prognostically unfavorable chromosomal aberration.

In the discovery cohort, absence of *IDH1* or *IDH2* mutation (11/98) was overrepresented in patients with distal 10q loss (7/11; five with whole chromosome 10 loss and two with distal 10q loss). After splitting the cohort by *IDH* status, a trend for distal 10q loss was observed; in the *IDH* mutant subgroup (n =87) the log rank test for loss of 10q (n =11) yielded a *P*-value of 0.077, and in the *IDH* wild-type subgroup (n =11), a similar *P*-value of 0.068 was yielded through the test for distal loss of 10q (n =7).

Co-deletion of 1p/19q was predominantly detected in LGGs with oligodendroglial histological features while loss of distal 10q was more frequently identified in astrocytic LGGs. However, there was no one-to-one relationship between histological features and these CNAs (Figure 2). Co-deletion of 1p/19q combined with distal 10q loss was observed in three LGGs of the discovery cohort and four LGGs of the validation cohort (all with oligodendroglial features) and none in the confirmation cohort. This limited number of patients does not allow for proper statistical survival analysis, but median survival of the patients in these cohorts combined (13.4 years) suggests that loss of 1p/19q and distal 10q counteracts overall survival.

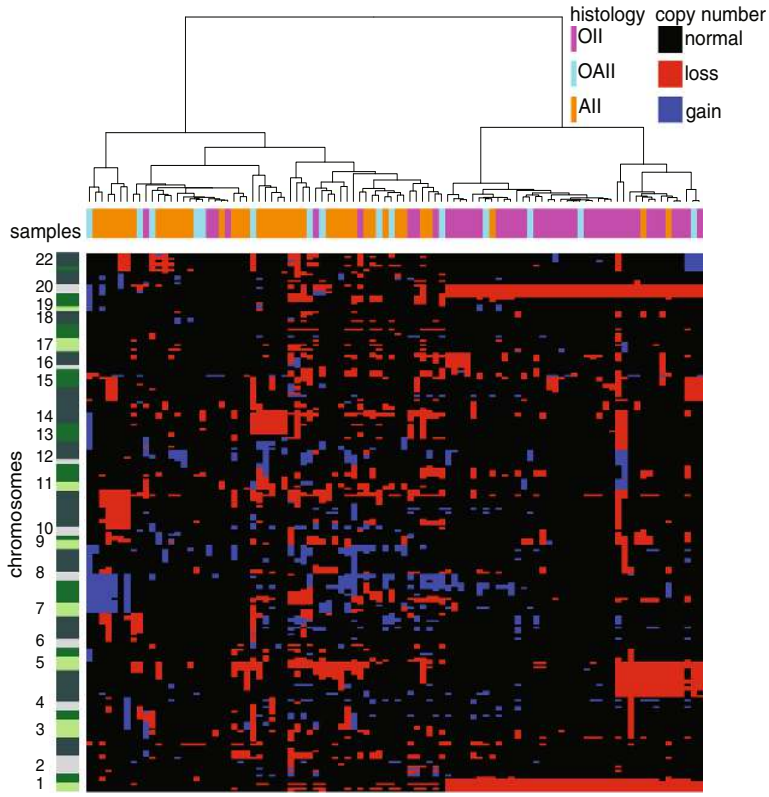

Figure 2 Unsupervised clustering of CNAs in the discovery cohort. Histological subtypes and patients are color-coded on the x-axis and chromosomes are ordered on the y-axis, 1 to 22 from bottom to top. Shades of green enable visualization of individual chromosomal arms, their size varying by the number of regions. Hence, a chromosomal arm with many breakpoints based on CNAs is depicted as larger compared with one with fewer breakpoints. Red, copy number loss; blue, copy number gain; black, no CNA. OII, oligodendroglioma; OAII, oligoastrocytoma; AII, astrocytoma.

Simultaneous testing of both CNAs classified LGG patients with a favorable (1p/19q co-deletion), unfavorable (distal 10q loss), or intermediate (both) prognosis in all three cohorts (Figure 3A,B,C). In the discovery cohort, hazard ratios of 1p/19q co-deletion without distal 10q loss and of distal 10q loss without 1p/19q co-deletion were 0.30 (95% confidence interval 0.15 to 0.58), and 2.91 (95% confidence interval 1.53 to 5.55), respectively (Table 3).

Intratumoral heterogeneity of CNAs in LGGs

In addition to the above-mentioned CNAs detected in single samples, we studied intratumoral heterogeneity by shallow WGS of multiple, spatially distinct regions obtained during the same surgery. Among other CNAs, distribution of 10q loss was assessed, illustrated for LGG240 in Figure 4 (more examples are provided in Additional file 2). In the original chromosomal profile of LGG240, marginal

Table 2 Prognostically unfavorable chromosomal regions of loss in diffuse low-grade gliomas

Chromosome	Start	End	Cytoband	Discovery cohort (%)	Validation cohort (%)	Discovery cohort (*P*-value)	Validation cohort (*P*-value)
9	24450001	28650000	9p21.3-21.1	21	7	0.039	1
10	112950001	135435000	10q25.2-qter	18	10	0.009	0.041
11	195001	14250000	11p15.5-15.2	13	15	0.0006	1
13	19500001	92550000	13q12.1-31.3	17	13	0.0001	1
22	34350001	51180000	22q12.3-13.33	11	8	0.000004	1

Frequency and *P*-value in discovery and validation cohorts calculated by log rank test and adjusted for multiple testing by Benjamini Hochberg and Holm Bonferroni, respectively. Positions according to GRCh37/hg19.

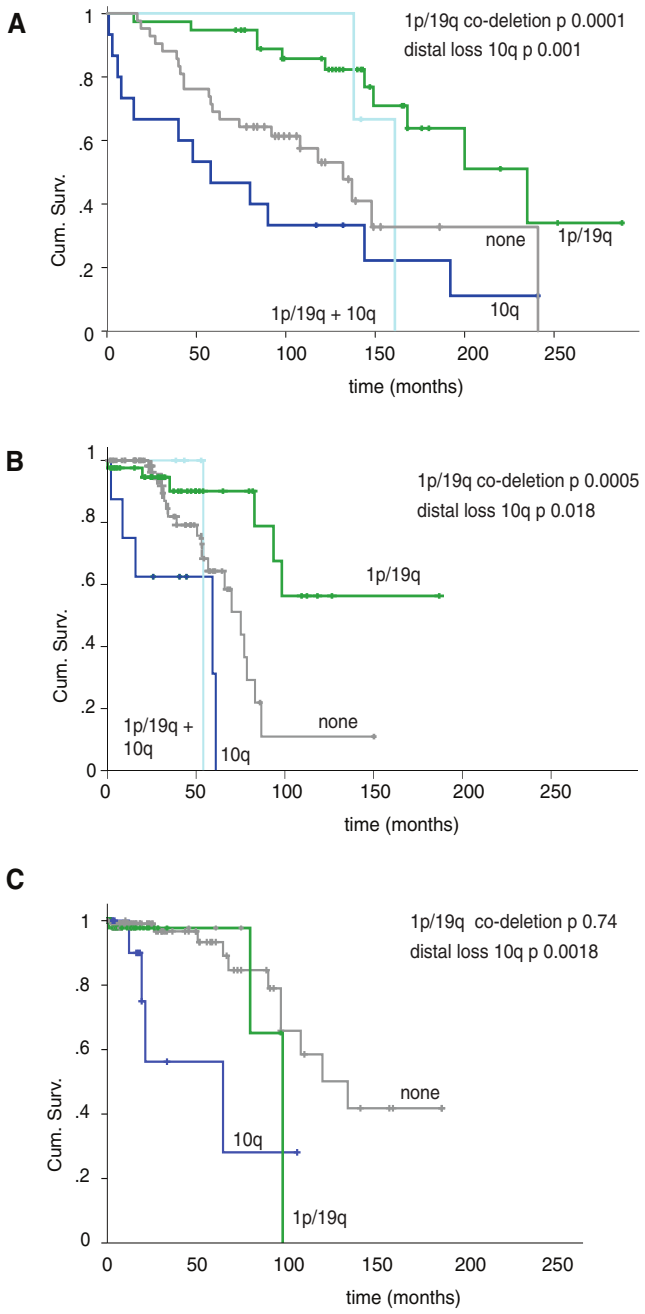

Figure 3 Kaplan Meier plots for distal 10q loss and 1p/19q co-deletion in (A) discovery, (B) validation and (C) confirmation cohorts. The dark blue line indicates loss of distal 10q without 1p/19q co-deletion versus the rest of the cohort (n =15, *P*-value =0.001 in **(A)**, n =8, *P*-value =0.018 in **(B)**, and n =14, *P*-value =0.0018 in **(C)**. The green line indicates 1p/19q co-deletion without distal loss of 10q (n =38, *P*-value =0.0001 in **(A)**, n =41, *P*-value =0.0005 in **(B)**, and n =47, *P*-value =0.74 in **(C)**. The light blue line indicates 10q loss and 1p/19q co-deletion (n =3, *P*-value =0.39 in **(A)**, n =4, *P*-value =0.94 in **(B)**, and n =0 in **(C)**. The grey line indicates neither 10q deletion nor 1p/19q co-deletion (n =42 in **(A)**, n =73 in **(B)**, and n =123 in **(C)**. The y-axis represents the fraction of patients alive, cumulative survival (Cum. Surv.), and the x-axis time in months. Censored patients are indicated with a vertical bar.

Table 3 Association of clinical and genetic parameters with overall survival in the discovery cohort

Parameter	n	*P*-value	HR
Age >50 years	23/98	0.187	1.57 (0.80-3.07)
Pre-operative KPS score <80	5/83	0.101	2.42 (0.79-7.05)
Pre-operative use of steroids	18/85	0.068	1.84 (0.99-3.58)
Pre-operative mass effect	37/71	0.0008	3.31 (1.61-6.73)
Pre-operative enhancement	35/79	0.031	2.16 (1.05-4.43)
Partial resection	69/90	0.029	2.85 (1.19-7.47)
Oligodendroglial histology	42/98	0.016	0.47 (0.25-0.87)
IDH1 or *IDH2* mutation	86/97	0.071	0.47 (0.21-1.07)
1p/19q co-deletion without 10q loss	41/98	0.0001	0.30 (0.15-0.58)
Loss of 10q without 1p/19q co-deletion	15/98	0.001	2.91 (1.53-5.55)

Results were determined using a log rank test. n, patients in subgroup compared with total number of patients with available data for each parameter; HR, hazard ratio (95% confidence interval). KPS, Karnofsky Performance Score.

deflection of 10q was observed (a smaller distance from the 0-line than for the losses in 1p, 4, or 19q (Figure 4A) [12]. Assuming clonality of the 1p/19q co-deletion [13], this difference in extent of copy number loss suggests that 10q loss would only be present in about 30 to 35% of the tumor cells (Figure S3 in Additional file 1).

To further delineate intratumoral heterogeneity of CNAs in this sample, the originally outlined area was divided into three sub-regions and an additional tumor region within the same paraffin section was included (Figure 4B). The 1p/19q co-deletion as well as chromosome 4 loss were present in all sub-regions and assumed to be clonally present. Losses of chromosomes 9, 10, 13, 15, 18 and gain of chromosome 11 were present in one or few sub-regions and assumed to be heterogeneously present (Figure 4C). To technically validate the intratumoral copy number heterogeneity observed in LGG240, array CGH was performed for all but one (insufficient amount of DNA) of the spatially distinct regions, which confirmed either clonality of 1p/19q and chromosome 4 losses or heterogeneity of all six chromosomally aberrant regions (Figure S4 in Additional file 1). Three additional samples with a clonal type of deflection and four with a marginal deflection of (distal) 10q were technically validated by array CGH (Figure S2 in Additional file 1). Intratumoral heterogeneity was detected in 15 out of 17 LGGs analyzed for this purpose; 68% of the CNAs (84/124) were not homogeneously present in spatially distinct regions obtained during the same surgery, such as loss of chromosomal arm 5q, chromosome 13 and gain of 11p (Figure 5A). Co-deletion of 1p and 19q was the only CNA that was consistently present in all spatially distinct regions of LGGs with this combination of CNAs; others, such as gain of chromosomal arm 7q, were most often, but not always, clonal. Loss of 10q was heterogeneously present in seven out of eight patients (Figure 5B). Histological variability did not correspond to the extent of heterogeneity.

Temporal evolution of CNAs in LGGs

Forty-seven out of 98 patients were subjected to a second surgery because of tumor progression. Of 20 patients, 24 recurrent tumors could be retrieved from medical archives. Almost 50% of CNAs (99/207) in the initial and paired recurrent tumors were shared and 15% (31/207) were uniquely detected in the initial tumor. A substantial proportion (37%, 77/207) of CNAs was uniquely identified in the recurrent tumor, such as loss of genomic regions at chromosomes 4, 10 and 15 (Figures 5C and 6). 1p/19q co-deletion was consistently identified in initial as well as recurrent tumors and there were no cases with new 1p/19q co-deletion. In four patients, *de novo* loss of 10q (including distal 10q losses) surfaced in the recurrence. In hindsight, marginal deflection of 10q was observed in the initial tumor of one of these four patients, and was not detected by the calling algorithm [12]. In two out of four patients with new loss of 10q a higher malignancy grade (WHO grade III or IV) had been assigned to the recurrent tumor (Figure 5D). In one of the three patients for which both spatially distinct regions of the initial tumor and recurrences were analyzed, subclonal 10q loss was present in one of the regions of the initial tumor, but undetectable in the recurrence (Additional file 3). In the other two patients, 10q loss was detected in both the initial and paired recurrent tumor.

Discussion

Intratumoral heterogeneity at the genomic level has been observed in numerous types of cancer, although its implications for treatment often remain undetermined. In the present study, archival LGG material and matched clinical data provide insight into spatial and temporal evolution of

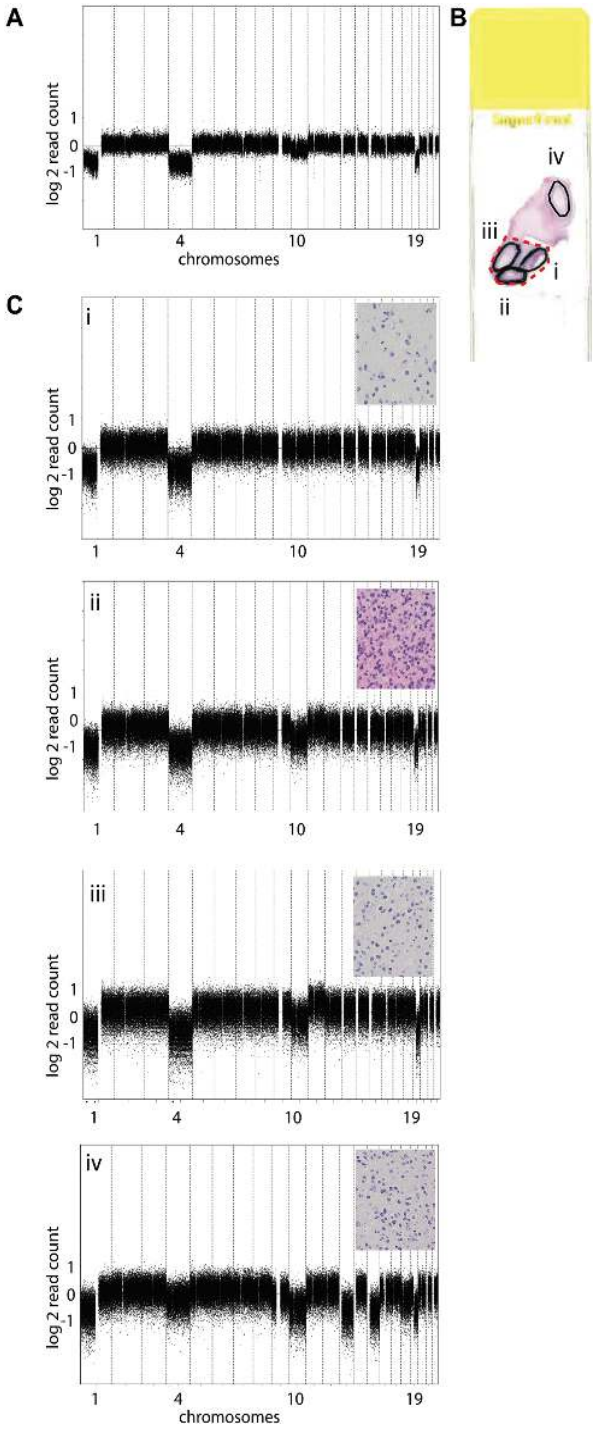

Figure 4 (See legend on next page.)

(See figure on previous page.)

Figure 4 Chromosomal copy number profiles for sample 240 demonstrating intratumoral copy number heterogeneity. (A) CNA profile of initial tumor, clonal 1p/19q co-deletion, loss of chromosome 4 and intermediate level of loss of chromosome 10. **(B)** Hematoxylin and eosin stained slide showing regions used for DNA isolation: the red dotted line corresponds to the region used for chromosomal profile of 4A and regions outlined with a solid black line (labeled i to iv) were used for the chromosomal profiles of 4C. **(C)** CNA profiles from four non-overlapping regions. Insets at the top right corner of each profile show histological features representative for individual regions. In all regions the histopathological diagnosis was LGG, although within a tumor the regions analyzed for spatial heterogeneity often showed some variation in microscopic features, such as cellularity and nuclear size and shape. (i) Clonal 1p/19q co-deletion and loss of chromosome 4; (ii) clonal 1p/19q co-deletion, loss of chromosome 4 and intermediate loss of chromosome 10; (iii) clonal 1p/19q co-deletion, loss of chromosome 4, intermediate loss of chromosome 10 and intermediate gain of chromosome 11; (iv) clonal 1p/19q co-deletion, intermediate loss of chromosomes 4, 10q, 13, 15 and 18. The y-axis represents normalized log2 sequence read counts per bin, and the x-axis represents 15 kb bins ordered by genomic position from chromosomes 1 to 22.

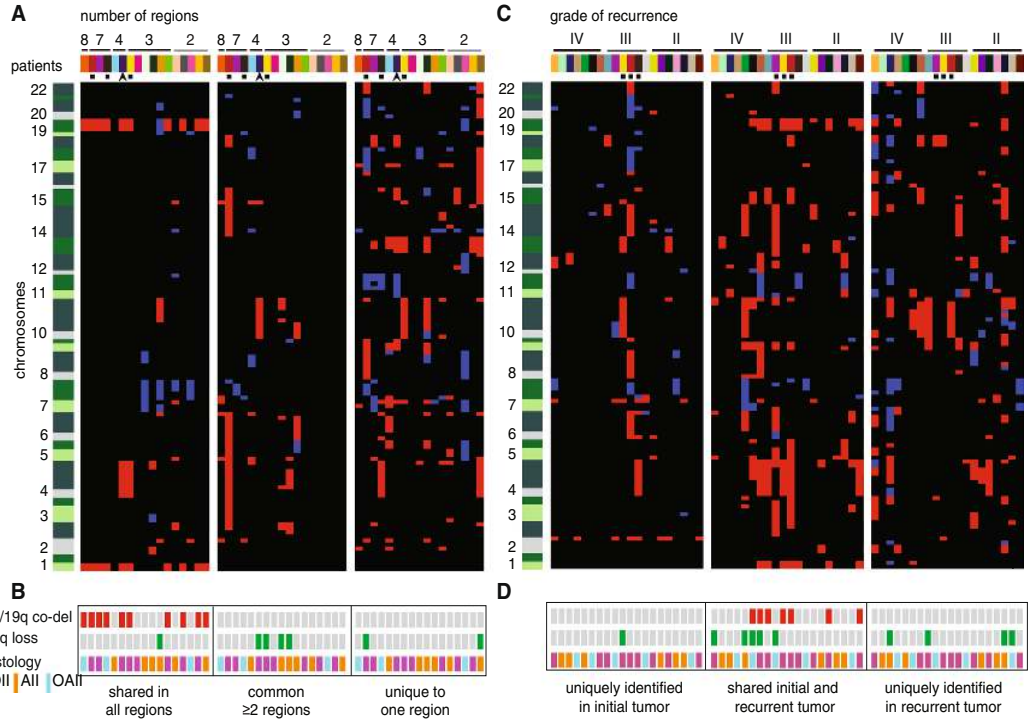

Figure 5 Spatial and temporal evolution of CNAs in LGGs and paired recurrent tumors. (A) CNAs in spatially distinct regions of LGGs of 17 patients. CNAs are categorized by detection in all regions (left panel), more than one region but not all regions (middle), or one region (right). Patients are ordered by the number of regions analyzed of each LGG from high to low. **(B)** Summary of prognostically relevant CNAs in spatially distinct regions and histology. No intratumoral heterogeneity was observed for 1p/19q co-deletion in any of the tumors, while distal 10q loss was often only detected in subclones. OII, oligodendroglioma; AII, astrocytoma; OAII, oligoastrocytoma. **(C)** CNAs in initial and paired recurrent tumors of 20 patients. CNAs are categorized by detection in initial tumor only (left panel), both initial and recurrence (middle) or detection uniquely in the recurrence (right). Patients are ordered by the histological malignancy grade of the recurrent tumor. **(D)** Summary of prognostically relevant CNAs in paired initial and recurrent tumors. 1p/19q co-deletion is stable over time, while distal 10q loss surfaces in recurrences, including two with a higher malignancy grade than LGG. Patients are color-coded on the x-axis and chromosomes are ordered on the y-axis, 1 to 22 from bottom to top. Shades of green enable visualization of individual chromosomal arms, their size varying by the number of regions. Hence, a chromosomal arm with many breakpoints based on CNAs is depicted as larger compared with one with fewer breakpoints. CNAs smaller than 5 Mbp were excluded from this figure. Red, copy number loss; blue, copy number gain; black, no CNA. The arrowhead indicates patient 240, the black squares three LGGs analyzed for both spatial and temporal evolution.

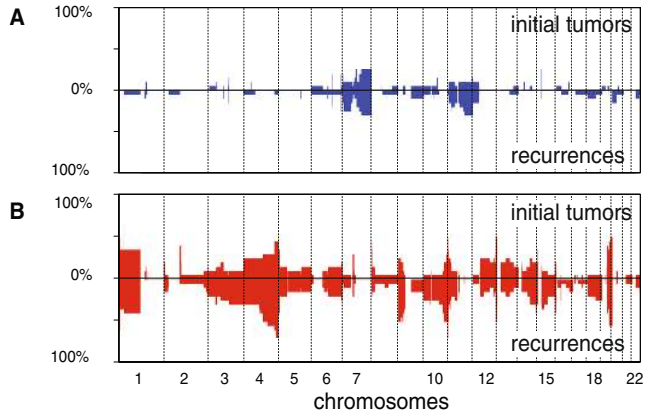

Figure 6 CNAs in initial and recurrent tumors. (A) Gains; **(B)** losses. The top of each graph shows the initial tumors, and the bottom the recurrences. Partial loss of chromosomal arm 4q, 9p and 10q were more frequently detected in recurrences. Bins are ordered by genomic position and from chromosomes 1 to 22 on the x-axis; percentages of cases showing CNAs are depicted on the y-axis.

prognostically relevant CNAs. All LGG samples collected for this study were included for copy number profiling by shallow WGS combined with a depth of coverage approach, yielding high quality data without technical drop-outs. This approach proved to be particularly beneficial for our study, since no matched normal DNA is required, which is a major advantage when analyzing long-term archived FFPE tumor samples. While shallow WGS cannot detect copy-neutral loss of heterozygosity or rearrangements, it is cost-effective, with a quality comparable to array CGH and applicable to DNA isolated from the FFPE samples. This allowed us to include samples that had been archived for over 30 years and collate a representative cohort, including LGG patients with long survival.

The relatively low incidence of LGGs and relatively long overall survival of patients necessitated this retrospective, multi-center approach. The observed variability in post-operative treatment can be attributed to the lack of a standard of treatment for these patients. Despite these variations, distal 10q loss (including whole chromosome losses) was significantly associated with an unfavorable prognosis in the discovery, validation and confirmation cohorts. Previously, some studies with smaller cohorts of specific histological subgroups of LGG have reported a correlation between 10q and survival [6,7]. Furthermore, a high prevalence of whole chromosome 10 loss and strong negative correlation with survival have been reported for grade III and IV gliomas [14-17]. Partial loss of 10q is much more frequently detected in histological grade II diffuse gliomas compared with grade III and IV gliomas. In each of the previously published studies the whole of chromosome 10 or the entire 10q arm was taken into account. Here we demonstrate that, different from higher grade gliomas, the distal end of 10q is frequently lost and associates with overall survival in three cohorts. Spatial as well as temporal analyses suggest that subclones with distal loss of 10q are involved in tumor progression, since the loss surfaces in paired recurrent tumors with a higher malignancy grade.

Identification of the genes and their proteins affected by CNAs may elucidate the biological underpinnings of their clinical relevance but can be challenging, especially when a genomic region is large. Only after many years were mutations in *CIC* and *FUBP1* associated with co-deletion of chromosomal arms 1p and 19q [18,19]. In total 148 genes are located on 10q25.2-qter, including *MGMT*, *DMBT1* and *ERCC6* [16,20,21], while the usual suspect, *PTEN*, is located more proximal to the centromere [16] and is preferentially lost in higher grade gliomas [4].

Based on our findings, we suggest that patients with an LGG should be simultaneously tested for both 1p/19q co-deletion and distal loss of 10q, since these two phenomena seem to have counteractive effects on survival. Introduction of heterogeneous CNAs, such as distal 10q loss, in daily clinical practice requires a robust diagnostic test. The well-known clonal features of 1p/19q co-deletion have been helpful to interpret these intermediate copy number levels in LGGs [13]. Analysis of multiple spatially distinct regions could reveal subclones. We currently favor genome-wide analysis, which visualizes chromosomes 1p, distal 10q and 19q simultaneously, and at the same time may provide insight into intratumoral heterogeneity within one region. Both extent of resection as well as the subclonal character of important markers for progression command alternative diagnostic procedures to assess their presence in a postsurgical situation, which may in

the future be offered through peripheral blood screening [4]. Meanwhile, detailed registration of the positions of samples from different regions within a tumor obtained during the same surgery may provide more accurate insight into biologically relevant topics such as the physical distance and direction of growth of tumor subclones as well as the overall extent of heterogeneity of a tumor [22].

Conclusions

Copy number analysis by shallow WGS is a robust approach for archival clinical LGG specimens. For a large proportion of LGG patients, analysis of CNAs with prognostic value may improve personalized timing of therapy. Thereby, loss of distal 10q without 1p/19q codeletion is indicative for urgent postoperative treatment, while in LGGs without loss of 10q and with 1p/19q codeletion a wait-and-scan policy should be considered. The subclonal character of whole or distal 10q loss in a subset of samples emphasizes the need for maximal extent of resection, illustrates that single sample diagnostics may be insufficient for LGG and favors future studies on genome-wide analysis of multiple spatially distinct samples to map tumor progression.

Materials and methods

Clinical data and sample collection for discovery, validation and confirmation cohorts

Approval for collection of clinical data and FFPE tumor samples for the 98 patients of the discovery cohort was obtained from the institutional review boards of all five Dutch hospitals, namely the Medical Ethical Committee (in Dutch: Medisch-Ethische Toetsingsingscommissie or METc) of the Academical Medical Center (AMC), the METc of the Isala klinieken in Zwolle, the METc of the VU University Medical Center (VUmc) in Amsterdam, the METc of the St Elisabeth Hospital in Tilburg, and the METc of the Arnhem - Nijmegen Region for samples from Radboud University Medical Center in Nijmegen (CMO). Experimental methods in this manuscript are in compliance with the Helsinki Declaration. Inclusion criteria and characteristics of the discovery cohort are summarized in Figure 1 and Table 1. Clinical features of the validation cohort, from a French hospital, can be found in Table 1; materials and methods are presented in more detail by Alentorn *et al.* [5]. For the confirmation cohort, copy number data of 531 lower grade glioma patients from the TCGA database were downloaded on 12 June 2014 via the Cancer Browser at UCSC [23]. Clinical data were available for 373 of these patients, including grade and overall survival; 184 of these samples were categorized as 'diffuse glioma histological grade 2' and selected as a confirmation cohort [10]. Presumably as a consequence of the fact that only fresh frozen samples were included in the TCGA cohort, a limited number of patients are contained in the dataset that had deceased during follow-up.

Laboratory techniques

Histological revision of samples in the discovery cohort was performed by two experienced neuropathologists (EA and PW). For all samples in the discovery cohort, including paired recurrent tumors, areas containing >60% tumor cells were outlined on hematoxylin and eosin stained slides, and tumor cell percentage estimated and registered for each sample (Additional files 2 and 3 and Table S1 in Additional file 4) and 10 adjacent sections were used for DNA isolation [24]. For the assessment of intratumoral heterogeneity, spatially distinct regions were selected based on histological variability and/or plain physical distance (Additional file 4). These samples were obtained either from one FFPE block, or individual blocks from the same surgery (Table S1 in Additional file 4). DNA (500 ng) was fragmented by sonication (Covaris™ S2, Woburn, MA, USA), and sequenced using a 50 bp single-read (50 bp SR) modus (Illumina TruSeq DNA-kit and HiSeq 2000, San Diego, CA, USA). The 100 bp paired-end (100 bp PE) sequencing modus and array CGH were used for comparison and technical validation. Array CGH was performed as described previously [25]. *IDH1* and *IDH2* mutation analysis was performed as described previously [5].

Statistical analysis

Copy number data from shallow WGS were analyzed using a novel Bioconductor script called QDNAseq [26]. QDNAseq infers copy numbers through depth of coverage by binning reads uniquely aligned to the human reference genome build GRCh37/hg19 with Burrow's Wheeler Alignment (BWA) [27]. PCR duplicates and reads with mapping qualities below 37 (highest value returned by BWA) were filtered. Copy numbers were inferred from the number of sequence reads per 15 kb bin. A simultaneous Loess correction for sequence mappability and GC content is applied within QDNAseq, which reduces noise of the copy number profiles, particularly for those with more degraded DNA. Problematic genome regions were furthermore filtered by applying our procedures to sequence data from the 1000 Genomes Project [28] to obtain a blacklist that eliminates problematic regions and the most common copy number variants of germ-line origin. Sequence data as well as all array CGH data have been uploaded to the European Genome-phenome Archive (EGA; accession number EGAS00001000643).

Calling of CNAs into discreet categories (normal, gain or loss) for the discovery and validation set was performed with the Bioconductor/R-package CGHcall [12]. A weighted hierarchical clustering of the CNAs

was performed using call probabilities to assess similarity of chromosomal profiles [29]. Association with survival was tested using a log rank test with significance estimated over 10,000 permutations. After discovery of regions of interest for survival, consecutive regions in the same chromosome with *P*-values <0.05 were fused together to final regions and the log rank test was repeated. Chromosomal regions that were still significant in the discovery set after multiple testing correction according to Benjamini-Hochberg were verified in the independent French validation cohort. Therefore, genomic coordinates were converted to the NCBI35/hg17 genome build using the UCSC liftOver tool [30] and *P*-values were calculated with the log rank test and adjusted with the more stringent Holm-Bonferroni method. The statistical significance of a CNA was calculated compared to the rest of the cohort not bearing this CNA - for example, samples with loss of distal 10q versus samples without loss of distal 10q.

Regions significant in both the discovery and validation cohorts were tested in the TCGA confirmation cohort for which CNA data were generated with Affymetrix SNP 6.0 arrays (Santa Clara, CA, USA). TCGA level 3 copy number data were publicly available at the time of download and mapped to NCBI36/hg18. These level 3 data involve beginning and end positions of chromosomal segments with deflection values, resulting from TCGA preprocessing (for level definitions and preprocessing see [31]). Segment values were converted to CNA discreet categories by setting thresholds whereby a log2 ratio of >0.20 is gain, < -0.23 is loss and all other values are normal copy number; these values correspond to 30% of the tumor cells with that CNA. Those segments overlapping for at least 90% of the 1p, 19q (excluding centromeres) or 10q25.2-qter region (corresponding to NCBI35/hg17: chr10: 112939991-135323881 and NCBI36/hg18: chr10:112939991-135284990) were taken into consideration. At this setting 14 patients had a distal 10q loss, of which 4 deceased during follow-up and no patients had both distal 10q and 1p/19q loss. *P*-value calculations were performed as described above without corrections since only one region was tested for confirmation. The threshold settings were selected based on the fact that a deletion in 30% of the tumor cells, as observed for the chromosome 10 loss in LGG sample 240 (Figure S3 in Additional file 1), should not be missed. All other threshold values for calling losses of these regions were stepwise tested as well as the percentage of overlap with the 10q25.2-qter region and are presented in Table S2 in Additional file 4. Significance for the 10q loss remained for many different settings, but the number of patients with this loss substantially decreased with widening margins for calling CNAs to lower than should be expected based on the discovery and validation cohorts.

To assess the presence of CNAs between spatially distinct regions and/or recurrences from the same patient, common regions were detected with CGHregions [32], and regions smaller than 5 Mbp were excluded. Clinical parameters were analyzed with log rank test. All reported *P*-values were two-sided, and <0.05 was considered statistically significant.

Data availability

Both array CGH and sequence data have been uploaded to the European Genome-phenome Archive (EGA; accession number EGAS00001000643).

Additional files

Additional file 1: Figure S1. Evaluation of shallow WGS for genome-wide copy number analysis for read lengths of sample LGG 284. Copy number profiles were produced by counting the number of uniquely mapped sequence tags per 15 kb bin for different settings. **(A)** Paired-end 100 (PE100), 100 bp reads from both ends. **(B)** Single-end 100 (SR100), 100 bp reads. **(C)** Single-end 50 (SR50), 50 bp reads from one end. **(D)** Technical validation by array CGH of the same sample. Y-axis, normalized log 2 sequence read counts per bin; x-axis, 15 kb bins ordered by genomic position from chromosomes 1 to22. **Figure S2.** Technical validation of shallow WGS by array CGH CNA analysis of eight LGGs tested on both platforms. The level of deflection of CNAs was comparable. Also, heterogeneously present CNAs with only marginal deflection could be reproduced; LGGs 168, 184 and 193 show clonal distal 10q loss, while LGGs 187, 189, 210, 211 and 224 show subclonal loss of 10q. **Figure S3.** Minimal tumor cell percentage required for detection of CNAs. Calculation of the percentage of tumor cells with a chromosome 10 deletion in LGG240. To simulate a limited tumor percentage, a virtual tumor-normal mixture experiment was performed with LGG240 (diffuse low-grade glioma) and NA18960 ((1000 Genomes Project Consortium). **Figure S4.** Technical validation of shallow WGS by array CGH. Technical validation of LGG240 (original) and three spatially distinct regions.

Additional file 2: Copy number profiles generated with shallow WGS of original LGGs of the discovery cohort and spatially distinct regions of this tumor obtained during the same surgery. A representative picture of histology (hematoxylin and eosin staining, original magnification × 200) is depicted in the top left corner of each profile.

Additional file 3: Copy number profiles generated with shallow WGS of initial LGGs and paired recurrent tumors. A representative picture of histology (hematoxylin and eosin staining, original magnification × 200) is depicted in the top left corner of each profile.

Additional file 4: Table S1. Overview of all samples of the discovery cohort, including paired recurrent tumors and spatially distinct regions. Columns show ID number, relative spatial distance (one FFPE block or individual FFPE block), histology, prognostically relevant CNAs, distal 10q loss, 1p/19q co-deletion, read count, sequence depth and tumor cell percentage as estimated by neuropathologists. **Table S2A-C.** Matrix with *P*-values for overall survival of the TCGA LGG (grade 2) confirmation cohort at various thresholds. Vertical rows: log2 ratio thresholds for calling CNA discrete categories (loss, normal or gain) from the log2 ratio level 3 data downloaded from the Cancer Browser at UCSC [23]. Horizontal rows: thresholds for overlap with the 10q25.2-qter region. Grey, non-significant settings; shades of beige, significant settings (darker indicates higher signiofícance); bold, threshold settings used for Figure 3. (A) *P*-values rounded to four decimal places for samples with or without a distal chromosome 10q loss. (B) Number of grade 2 LGG patients with clinical data (n =184) from the lower grade glioma TCGA dataset with a distal 10q deletion. (C) Number of grade 2 LGG patients (n =184) from the lower grade glioma TCGA dataset with a distal 10q deletion that had deceased during follow-up. Overall survival is significant for various settings of distal 10q loss. The number of patients with distal 10q loss substantially decreased with widening margins for calling CNAs, to lower than should be expected based on the French and Dutch cohorts.

Abbreviations
bp: base pair; CGH: comparative genomic hybridization; CNA: copy number aberration; FFPE: formalin-fixed paraffin-embedded; LGG: diffuse low-grade glioma; PCR: polymerase chain reaction; TCGA: The Cancer Genome Atlas; WGS: whole genome sequencing.

Competing interests
AI: Novartis, Hoffmann-La Roche, IntselChimos and BetaInnov, no conflict of interest related to the present work. KHX: LOC network (primary CNS lymphoma) supported by the INCa (National Institute of Cancer). AA: Obra Social 'la Caixa' and ARTC. The other authors report no competing interests.

Authors' contributions
HFT: study conceptualization and design, sample and clinical data collection, DNA isolation, statistical analysis and manuscript writing. IS: study conceptualization and design, sequence data generation and interpretation, bioinformatics analysis, manuscript writing. HFE: DNA isolation, sequence library preparations, and manuscript review. AA: validation cohort, data collection, analysis, interpretation and manuscript review. DS and MC: sequence data generation and bioinformatics, manuscript review. RF: sample collection, histological analysis and manuscript review. AMG: clinical data collection and manuscript review. GB: clinical data collection and manuscript review. WAB: clinical data collection and manuscript review. GAM: histological analysis and manuscript review. MH: sample collection, histological analysis and manuscript review. AI: validation cohort, clinical data collection, study design, data interpretation and manuscript review. KHX: validation cohort clinical data collection and manuscript review. KM: histological analysis and manuscript review. RGV: confirmation cohort analysis and manuscript review. PV: study conceptualization, histological analysis and manuscript review. MAW: statistical analysis, bioinformatics and manuscript review. JJH: study conceptualization and design, manuscript writing. EA: histological analysis and revision, manuscript review. JCR: study conceptualization and design, clinical data collection and analysis, manuscript writing. PW: study conceptualization and design, sample collection, histological analysis and revision, manuscript writing. BY: study conceptualization and design, sequence data generation, sample collection, bioinformatics, manuscript writing. All authors read and approved the final manuscript.

Authors' information
The present study was performed by a wide variety of professionals. HFT is a PhD student in neuro-oncology and resident in neurology; IS, DS and MC are PhD students specialized in bioinformatics; HE is a research technician specialized in DNA isolation from FFPE, microarrays and next generation sequencing; AA, AMG, AI, KHX, JJH and JCR are neuro-oncologists providing clinical care for brain tumor patients and are involved in several research projects in this field; GB and WAB are neurosurgeons that provide daily clinical care for brain tumor patients; RF, MH, EA, GAM, PV and PW are pathologists with a special interest in neuro-oncology; MW is a mathematician and statistical expert in genome-wide data analysis; RGV and BY are molecular biologists specialized in tumor genome analysis.

Acknowledgments
This work was supported by grants from the Dutch Cancer Society (KWF kankerbestrijding, grant nr 2009-4470), foundation STOPHersentumoren.nl and Edli foundation. The validation cohort is part of the national program Cartes d'Identité des Tumeurs® (CIT). The research leading to these results received funding from the program 'Investissements d'avenir' ANR-10-IAIHU-06.

Author details
[1]Department of Pathology, VU University Medical Center, 1007 MB Amsterdam, The Netherlands. [2]Department of Neurology, VU University Medical Center, 1007 MB Amsterdam, The Netherlands. [3]Department of Pathology, Haartman Institute and HUSLAB, University of Helsinki and Helsinki University Central Hospital, 00014 Helsinki, Finland. [4]Université Pierre et Marie Curie-Paris 6 Centre de Recherche de l'Institut du Cerveau et de la Moelle Epinière (CRICM), 75013 Paris, France. [5]Inserm U975, 75013 Paris, France. [6]Centre National de la Recherche Scientifique (CNRS), 75013 Paris, France. [7]AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Department of Neurology 2-Mazarin, 75013 Paris, France. [8]Department of Pathology, Elisabeth Hospital Tilburg, 5022 GC Tilburg, The Netherlands. [9]Department of Neurology, Radboud University Medical Center, 6525 GANijmegen, The Netherlands. [10]Department of Neurological Surgery, Elisabeth Hospital Tilburg, 5022 GC Tilburg, The Netherlands. [11]Department of Neurological Surgery, Isala, 8011 JW Zwolle, The Netherlands. [12]Department of Pathology, Isala, 8011 JW Zwolle, The Netherlands. [13]AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Department of Neuropathology, 2-Mazarin, 75013 Paris, France. [14]Departments of Genomic Medicine, University of Texas, MD Anderson Cancer Center, Houston, TX 77054, USA. [15]Department of Bioinformatics and Computational Biology, University of Texas, MD Anderson Cancer Center, Houston, TX 77230, USA. [16]Department of Epidemiology and Biostatistics, VU University Medical Center, 1007 MB Amsterdam, The Netherlands. [17]Department of Mathematics, VU University, 1081 HV Amsterdam, The Netherlands. [18]Department of Pathology, Academical Medical Center, 1105 AZ Amsterdam, The Netherlands. [19]Department of Neurology, Academical Medical Center, 1105 AZ Amsterdam, The Netherlands. [20]Department of Pathology, Radboud University Medical Center, 6525 GA Nijmegen, The Netherlands.

Received: 11 July 2014 Accepted: 15 September 2014
Published online: 23 September 2014

References

1. Claus EB, Black PM: **Survival rates and patterns of care for patients diagnosed with supratentorial low-grade gliomas: data from the SEER program, 1973-2001.** *Cancer* 2006, **106:**1358–1363.
2. Douw L, Klein M, Fagel SS, van den Heuvel J, Taphoorn MJ, Aaronson NK, Postma TJ, Vandertop WP, Mooij JJ, Boerman RH, Beute GN, Sluimer JD, Slotman BJ, Reijneveld JC, Heimans JJ: **Cognitive and radiological effects of radiotherapy in patients with low-grade glioma: long-term follow-up.** *Lancet Neurol* 2009, **8:**810–818.
3. Erdem-Eraslan L, Gravendeel LA, de Rooi J, Eilers PH, Idbaih A, Spliet WG, den Dunnen WF, Teepen JL, Wesseling P, Sillevis Smit PA, Kros JM, Gorlia T, van den Bent MJ, French PJ: **Intrinsic molecular subtypes of glioma are prognostic and predict benefit from adjuvant procarbazine, lomustine, and vincristine chemotherapy in combination with other prognostic factors in anaplastic oligodendroglial brain tumors: a report from EORTC study 26951.** *J Clin Oncol* 2013, **31:**328–336.
4. van Thuijl HF, Ylstra B, Wurdinger T, van Nieuwenhuizen D, Heimans JJ, Wesseling P, Reijneveld JC: **Genetics and pharmacogenomics of diffuse gliomas.** *Pharmacol Ther* 2013, **137:**78–88.
5. Alentorn A, van Thuijl HF, Marie Y, Alshehhi H, Carpentier C, Boisselier B, Laigle-Donadey F, Mokhtari K, Scheinin I, Wesseling P, Ylstra B, Capelle L, Hoang-Xuan K, Delattre JY, Reijneveld JC, Idbaih A: **Clinical value of chromosome arms 19q and 11p losses in low-grade gliomas.** *Neuro Oncol* 2014, **16:**400–408.
6. Bissola L, Eoli M, Pollo B, Merciai BM, Silvani A, Salsano E, Maccagnano C, Bruzzone MG, Fruhman Conti AM, Solero CL, Giombini S, Broggi G, Boiardi A, Finocchiaro G: **Association of chromosome 10 losses and negative prognosis in oligoastrocytomas.** *Ann Neurol* 2002, **52:**842–845.
7. Houillier C, Mokhtari K, Carpentier C, Criniere E, Marie Y, Rousseau A, Kaloshi G, Dehais C, Laffaire J, Laigle-Donadey F, Hoang-Xuan K, Sanson M, Delattre JY: **Chromosome 9p and 10q losses predict unfavorable outcome in low-grade gliomas.** *Neuro Oncol* 2010, **12:**2–6.
8. Snuderl M, Fazlollahi L, Le LP, Nitta M, Zhelyazkova BH, Davidson CJ, Akhavanfard S, Cahill DP, Aldape KD, Betensky RA, Louis DN, Iafrate AJ: **Mosaic amplification of multiple receptor tyrosine kinase genes in glioblastoma.** *Cancer Cell* 2011, **20:**810–817.
9. Johnson BE, Mazor T, Hong C, Barnes M, Aihara K, McLean CY, Fouse SD, Yamamoto S, Ueda H, Tatsuno K, Asthana S, Jalbert LE, Nelson SJ, Bollen AW, Gustafson WC, Charron E, Weiss WA, Smirnov IV, Song JS, Olshen AB, Cha S, Zhao Y, Moore RA, Mungall AJ, Jones SJ, Hirst M, Marra MA, Saito N, Aburatani H, Mukasa A, *et al*: **Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma.** *Science* 2014, **343:**189–193.
10. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu CJ, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla SA, Ciriello G, Yung WK, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner DD, van Meir EG, Prados M, Sloan A, Black KL, *et al*: **The somatic genomic landscape of glioblastoma.** *Cell* 2013, **155:**462–477.

11. Idbaih A, Criniere E, Ligon KL, Delattre O, Delattre JY: **Array-based genomics in glioma research.** *Brain Pathol* 2010, **20:**28–38.
12. van de Wiel MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, Ylstra B: **CGHcall: calling aberrations for array CGH tumor profiles.** *Bioinformatics* 2007, **23:**892–894.
13. Jeuken JW, Sijben A, Bleeker FE, Boots-Sprenger SH, Rijntjes J, Gijtenbeek JM, Mueller W, Wesseling P: **The nature and timing of specific copy number changes in the course of molecular progression in diffuse gliomas: further elucidation of their genetic 'life story'.** *Brain Pathol* 2011, **21:**308–320.
14. Horbinski C, Nikiforova MN, Hobbs J, Bortoluzzi S, Cieply K, Dacic S, Hamilton RL: **The importance of 10q status in an outcomes-based comparison between 1p/19q fluorescence in situ hybridization and polymerase chain reaction-based microsatellite loss of heterozygosity analysis of oligodendrogliomas.** *J Neuropathol Exp Neurol* 2012, **71:**73–82.
15. Ramirez C, Bowman C, Maurage CA, Dubois F, Blond S, Porchet N, Escande F: **Loss of 1p, 19q, and 10q heterozygosity prospectively predicts prognosis of oligodendroglial tumors–towards individualized tumor treatment?** *Neuro Oncol* 2010, **12:**490–499.
16. Sasaki H, Zlatescu MC, Betensky RA, Ino Y, Cairncross JG, Louis DN: **PTEN is a target of chromosome 10q loss in anaplastic oligodendrogliomas and PTEN alterations are associated with poor prognosis.** *Am J Pathol* 2001, **159:**359–367.
17. Thiessen B, Maguire JA, McNeil K, Huntsman D, Martin MA, Horsman D: **Loss of heterozygosity for loci on chromosome arms 1p and 10q in oligodendroglial tumors: relationship to outcome and chemosensitivity.** *J Neurooncol* 2003, **64:**271–278.
18. Bettegowda C, Agrawal N, Jiao Y, Sausen M, Wood LD, Hruban RH, Rodriguez FJ, Cahill DP, McLendon R, Riggins G, Velculescu VE, Oba-Shinjo SM, Marie SK, Vogelstein B, Bigner D, Yan H, Papadopoulos N, Kinzler KW: **Mutations in CIC and FUBP1 contribute to human oligodendroglioma.** *Science* 2011, **333:**1453–1455.
19. Yip S, Butterfield YS, Morozova O, Chittaranjan S, Blough MD, An J, Birol I, Chesnelong C, Chiu R, Chuah E, Corbett R, Docking R, Firme M, Hirst M, Jackman S, Karsan A, Li H, Louis DN, Maslova A, Moore R, Moradian A, Mungall KL, Perizzolo M, Qian J, Roldan G, Smith EE, Tamura-Wells J, Thiessen N, Varhol R, Weiss S, *et al.*: **Concurrent CIC mutations, IDH mutations, and 1p/19q loss distinguish oligodendrogliomas from other cancers.** *J Pathol* 2012, **226:**7–16.
20. Boulay JL, Ionescu MC, Sivasankaran B, Labuhn M, Dolder-Schlienger B, Taylor E, Morin P Jr, Hemmings BA, Lino MM, Jones G, Maier D, Merlo A: **The 10q25.3-26.1 G protein-coupled receptor gene GPR26 is epigenetically silenced in human gliomas.** *Int J Oncol* 2009, **35:**1123–1131.
21. Ramalho-Carvalho J, Pires M, Lisboa S, Graca I, Rocha P, Barros-Silva JD, Sawa-Bordalo J, Mauricio J, Resende M, Teixeira MR, Honavar M, Henrique R, Jeronimo C: **Altered expression of MGMT in high-grade gliomas results from the combined effect of epigenetic and genetic aberrations.** *PLoS One* 2013, **8:**e58206.
22. Sottoriva A, Spiteri I, Piccirillo SG, Touloumis A, Collins VP, Marioni JC, Curtis C, Watts C, Tavare S: **Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics.** *Proc Natl Acad Sci U S A* 2013, **110:**4009–4014.
23. Zhu J, Sanborn JZ, Benz S, Szeto C, Hsu F, Kuhn RM, Karolchnik D, Archi J, Lenburg ME, Esserman LJ, Kent WJ, Haussler D, Wangt T: **The UCSC Cancer Genomics Browser.** *Nat Methods* 2009, **6:**239–240.
24. van Essen HF, Ylstra B: **High-resolution copy number profiling by array CGH using DNA isolated from formalin-fixed, paraffin-embedded tissues.** *Methods Mol Biol* 2012, **838:**329–341.
25. Krijgsman O, Israeli D, van Essen HF, Eijk PP, Berens ML, Mellink CH, Nieuwint AW, Weiss MM, Steenbergen RD, Meijer GA, Ylstra B: **Detection limits of DNA copy number alterations in heterogeneous cell populations.** *Cell Oncol (Dordr)* 2013, **36:**27–36.
26. Scheinin I, Sie D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF, van Essen HF, Eijk PP, Rustenburg F, Meijer GA, Reijneveld JC, Wesseling P, Pinkel D, Albertson DG, Ylstra B: **DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly.** *Genome Res* 2014, gr.175141.114.
27. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754–1760.
28. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491:**56–65.
29. van Wieringen WN, van de Wiel MA, Ylstra B: **Weighted clustering of called array CGH data.** *Biostatistics* 2008, **9:**484–500.
30. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, Hillman-Jackson J, Kuhn RM, Pedersen JS, Pohl A, Raney BJ, Rosenbloom KR, Siepel A, Smith KE, Sugnet CW, Sultan-Qurraie A, Thomas DJ, Trumbower H, Weber RJ, Weirauch M, Zweig AS, Haussler D, Kent WJ: **The UCSC Genome Browser Database: Update 2006.** *Nucleids Acids Res* 2006, **1:**D590–8.
31. **The Cancer Genome Atlas Data Portal** [https://tcga-data.nci.nih.gov/tcga]
32. van de Wiel MA, Wieringen WN: **CGHregions: dimension reduction for array CGH data with minimal information loss.** *Cancer Inform* 2007, **3:**55–63.

doi:10.1186/s13059-014-0471-6
Cite this article as: van Thuijl *et al.*: Spatial and temporal evolution of distal 10q deletion, a prognostically unfavorable event in diffuse low-grade gliomas. *Genome Biology* 2014 **15**:471.

Chapter 6

Summary and discussion

Summary of the original publications

The previous four chapters present original, peer-reviewed publications that describe three bioinformatic solutions for the analysis of CNAs in cancer, and the application of one of them on survival analysis of diffuse low-grade glioma patients. This chapter starts with a brief summary of each original publication, and subsequently discusses them together from today's perspective.

CanGEM database for CNAs in cancer

Chapter 2 (Scheinin et al., 2008) describes a database developed for array CGH data from cancer samples. The CanGEM (Cancer GEnome Mine) database is compliant with the MIAME standard (Brazma et al., 2001) and accessible publicly on the Internet. The areas where CanGEM extends upon existing microarray databases, such as GEO (Barrett et al., 2013) and ArrayExpress (Kolesnikov et al., 2015), is in sample meta-data and microarray *copy number analysis*, and the subsequent querying functionality.

The reason we decided to develop a new database was that while the existing microarray databases fulfilled an important role as data repositories for published articles, they were not designed to be approached from the other direction: the data itself. The primary use of existing databases was to allow data sets to be retrieved with an accession number listed in a published article. What we envisioned with CanGEM was a resource designed to be queried based on clinical variables or processed microarray data. A solution that was designed to answer questions such as "show me all samples with an *EGFR* amplification", or "all breast cancer cases diagnosed before age 30". Existing databases were not build for these types of tasks, and that is why we decided to develop CanGEM.

Clinical data

Since the existing microarray databases house data from different types of experiments on various different species, little structure is imposed on sample meta-data. It is typically collected in the form of free-text fields, where users enter what is in their opinion the most relevant description. Therefore, it can be difficult, or even impossible, to perform complex queries based on sample meta-data.

Since the CanGEM database is focused on CNAs in cancer, it uses relevant controlled vocabularies for sample annotations. These include chapter II (Neoplasms) of the International Statistical Classification of Diseases and Related Health Problems (10th Revision; ICD-10) and International Classification of Disease for Oncology (3rd Edition; ICD-O-3) of the World Health Organization (WHO), the TNM Classification (Tumor, lymph Nodes, Metastasis), and six of the ten categories in the eVOC Ontology (Kelso et al., 2003). Information is also stored on patient sex, surgery outcome, survival, cause of death, tumor size, and exposure to environmental factors. All of these attributes can be used to query for samples across studies in the CanGEM database.

Copy number analysis of microarray data

Another defining factor for the existing general-purpose microarray databases is that they contain data from different types of microarray experiments, from mRNA and miRNA expression to DNA copy number and methylation. This makes it a challenge to implement querying functionality based on the microarray data itself.

As CanGEM is focused on chromosomal aberrations, *copy number analysis* of the raw microarray data is performed with *preprocessing*, *segmentation*, and *calling* (Figure 1.2). This allows users to perform queries for

specific CNAs, such as the example question mentioned above. The *copy number analysis* also allows overall aberration frequencies to be calculated over desired data sets, which can be defined as the result sets of search queries, or constructed manually by combining individual studies or samples within them.

Due to the evolving picture of the human reference genome, different generations of microarray platforms have been mapped to different builds of the reference genome. To make them comparable, microarray platforms added to CanGEM are re-annotated through a Megablast analysis of the DNA sequences of the array elements (Zhang et al., 2000). This is performed separately for each genome build. In addition to simplifying comparison and combination of results from different array generations, this type of remapping has also been shown to lead to improved accuracy over mappings provided by array manufacturers (Elo et al., 2005).

The *copy number analysis* pipeline of CanGEM consists of LOESS normalization (Smyth and Speed, 2003) followed by *segmentation* and *calling*, which were originally performed with CGH-Explorer (Lingjaerde et al., 2005). After publication of the original article that describes the CanGEM database, the pipeline was updated to use DNAcopy (Venkatraman and Olshen, 2007) for *segmentation* and CGHcall (van de Wiel et al., 2007) for *calling*. Since different platforms have array elements that target different positions in the genome, the final step of the processing pipeline is to define a copy number *call* for every known human gene. The list of their positions was retrieved from the Ensembl database (Yates et al., 2016). For genes whose positions in the genome overlaps with positions of elements on the array, those overlapping elements are used to derive the gene-specific copy number *call*. For genes with no overlapping array elements, the last preceding and first tailing element are used. This allows results measured with different microarray platforms to be integrated and queried together.

Sample size calculations with CGHpower

Chapter 3 (Scheinin et al., 2010) describes a dedicated solution for sample size calculations in the context of genome-wide copy number experiments that compare two groups of cancer samples. It estimates average power as a function of sample size, and uses a pilot data set to estimate parameters. It is a combination of a *copy number analysis* pipeline and a statistical framework previously developed for sample size calculations in the context of mRNA expression arrays (Ferreira and Zwinderman, 2006a).

The question how many samples are statistically required affects every experiment that aims to compare two groups of patients. It is a crucial factor that affects not only the scientific results of the experiments, but also has a role in justifying the amount of funding sought with research grants. While there were existing solutions to other such situations, no tool had been developed for the specific context of cancer CNAs. Methods for sample size calculations usually either ask the user to specify relevant parameters (such as the effect size between groups, variance, and desired levels of statistical significance and power) or estimate these parameters from existing data. Because issues such as varying cellularity, ploidy, and tumor heterogeneity can affect observed effect sizes, they can be difficult to estimate without prior measurement data from the tumor type in question. Therefore, we chose to develop a dedicated solution that used a pilot data set to perform sample size calculations for experiments that compare CNAs between two groups of patients.

Copy number analysis and power calculations

In the *copy number analysis* worklow (Figure 1.6), there are a number of points that could be used as the basis for sample size

calculations. After evaluation of alternative strategies, the approach we chose was to use a framework developed for sample size calculations for mRNA expression arrays (Ferreira and Zwinderman, 2006a), and combine it with a *copy number analysis* workflow that consists of *segmentation* (Venkatraman and Olshen, 2007), *calling* (van de Wiel et al., 2007), *regioning* (van de Wiel and van Wieringen, 2007), and finally representing each region with its median $\log_2$-ratio.

For each region, t-statistics and p-values from the normal distribution are calculated between the two groups. Some of these regions are expected to exhibit a true difference between the two groups, while others are not. The p-values for the regions therefore come from two separate distributions. For the ones with no real difference, this distribution is the uniform. For the ones with a difference, the distribution is unknown. This unknown distribution is estimated with two estimators. These estimators depend on another unknown parameter, which is the proportion of regions with no real difference between groups. This proportion is estimated by minimizing the distance between the two estimators for the unknown p-value distribution. Once these estimates and the limiting density of effect sizes have been calculated, an adaptive version of the Benjamini-Hochberg method for multiple testing is used to estimate average power as a function of sample size (Ferreira and Zwinderman, 2006b).

Diagnostic plots

The accuracy of power estimates produced by CGHpower depend on the accuracy of model parameters, which are estimated from a pilot data set. If parameter estimation performs poorly, validity of the power estimates will also be questionable. The program output contains a set of diagnostic plots that can be used to assess the quality of parameter estimation.

Performance of CGHpower was evaluated with both simulated and real data sets. Some cases were found to perform well, but in a number of them parameter estimation showed poor performance. This can happen when model assumptions are not fulfilled. In particular, the proportion of regions with no real differences between groups has to be substantially less than one. In other words, if only a very small number of regions (such as 1 %) show a true difference between the two groups, there might be too few data points available for estimation. While it is clear the performance of the program is questionable for some data sets, the diagnostic plots allow its reliability to be assessed.

Copy number preprocessing with QDNAseq

Chapter 4 (Scheinin et al., 2014) describes QDNAseq, a DOC preprocessing method to detect CNAs from NGS data. It improves upon existing methods by 1) introducing a simultaneous correction for GC content and mappability, and 2) filtering out spurious regions in the genome. It does not require use of a control sample, can be used for a wide range of sequence coverage, and performs well with FFPE samples.

When we started with this project, the rapid technological improvements and decreasing costs of NGS were turning it into a more appealing alternative to microarrays. While the technology showed a lot of promise, data processing workflows to detect CNAs were limited. False positive CNAs were observed frequently and at recurrent locations, and the published methods were often difficult to fit in full analytical workflows of larger data sets.

Correction to read counts and identification of problematic regions in the genome

GC content and mappability are both factors that are known to affect observed sequence coverage in NGS experiments. DOC methods therefore usually include a correction for GC content, and a correction and/or filtering for mappability. QDNAseq was the first method

to introduce a simultaneous LOESS correction for both GC content and mappability. In case there was any interaction between the two biases, a simultaneous correction could lead to improved performance. Also, it is an approach that can be extended to further dimensions, if additional sources of systematic bias are discovered in the future.

The human genome includes regions that are problematic to sequence and characterize with current methods. Long stretches of highly repetitive sequence are hard to assemble into an accurate reference genome. These regions are especially pronounced near the centromeres and telomeres, but can also be found elsewhere in the genome. If left unfiltered, these regions result in spurious *calls* and make results more difficult to interpret. ENCODE Project Consortium et al. (2012) recommends that these regions are filtered from analyses, and has published a blacklist of such regions.

QDNAseq introduces a novel blacklist, which is based on its simultaneous LOESS correction for GC content and mappability, and a control data set from the 1000 Genomes project (1000 Genomes Project Consortium et al., 2012). During the LOESS correction of each sample, a residual was defined for each bin. Median residuals were then calculated across the data set, and bins with the absolute value of median residuals higher than 4.0 standard deviations were included in the blacklist. Both blacklists are optional and can be used to filter bins from the analysis.

Performance evaluation

Performance of QDNAseq was evaluated in comparison to an existing DOC preprocessing method, array CGH, and also to theoretical expectations. Compared to array CGH, both NGS methods were shown to have higher signal-to-noise ratios. Through a comparison to a method with separate corrections for GC content and mappability (Boeva et al., 2011), we showed the simultaneous correction to be always at least as good, and sometimes superior. Furthermore, the noise of the *preprocessed* data (as measured with a mean-scaled and 0.1%-trimmed first-order estimate of variance) was shown to be very close to the theoretical statistical limit imposed by read counting.

CNAs in low-grade gliomas

Chapter 5 (van Thuijl et al., 2014) utilizes low-coverage WGS and the QDNAseq method described in Chapter 4 (Scheinin et al., 2014) to identify CNAs in archival material from 98 patients with diffuse low-grade glioma (LGG). Detected CNAs are used to evaluate their association with survival, and also to study intratumoral heterogeneity and temporal evolution of LGGs.

LGGs are brain tumors that grow relatively slow. Patients can live up to 30 years with the disease, but in some cases survival can be as short as two years (Claus and Black, 2006). Post-operative radiotherapy is effective in limiting tumor growth, but since its side effects can be substantial, an accurate characterization of the disease is crucial for appropriate decisions on post-operative treatment (Douw et al., 2009). CNA markers associated with survival could thus be highly valuable.

Associations between CNAs and survival

LGGs commonly exhibit a 1p/19q co-deletion, which is well known to be associated with a less aggressive disease and better prognosis (Erdem-Eraslan et al., 2013; van Thuijl et al., 2012). In addition to this co-deletion, other CNAs that were found to be statistically significantly associated with survival included losses in 10q, 11p, 13q, and 22q, which were all associated with poor prognosis. Of these, the 10q loss was validated in an independent French cohort of 126 patients (Alentorn et al., 2014), and confirmed also in a set of 184 patients from The Cancer Genome Atlas (TCGA) (Brennan et al., 2013). The size of the loss varied from the whole chromosome 10 to a 22.5 Mbp minimal common

region spanning from 10q25.5 to the end of the long arm. This region contains 148 genes, including for example *MGMT*, *DMBT1*, and *ERCC6*, but not *PTEN*.

Evolving picture of glioma classification

When we started the low-grade glioma project, the official classification of LGGs included three histological subtypes: astrocytomas, oligodendrogliomas, and the mixed phenotype oligoastrocytoma. The 1p/19q co-deletion (Reifenberger et al., 1994) was known to be common among oligodendrogliomas, but there were also cases of oligodendrogliomas without the co-deletion, and cases of co-deletions in oligoastrocytomas and even astrocytomas. Our set of 98 patients reflected this, and contained cases of all three histological subtypes and both with and without the co-deletion.

The understanding of tumorigenesis of gliomas has since improved rapidly, which has resulted in a new classification system, now based not only on histopathological features but also on molecular markers (Louis et al., 2016). With the use of molecular markers, it should now be possible to define all cases previously diagnosed as the mixed oligoastrocytoma subtype as either astrocytomas or oligodendrogliomas. The defining features of oligodendrogliomas are the presence of both the 1p/19q co-deletion and a mutation in either *IDH1* or *IDH2* genes (Yan et al., 2009). The new classification system shows stronger association between disease subtype and prognosis, and provides a better basis for treatment decisions (Wesseling et al., 2015).

Consistent with the role of the 1p/19q co-deletion as a hallmark feature, are our observations regarding intratumoral heterogeneity and temporal evolution. When possible, intratumoral heterogeneity was assessed by isolating DNA from either spatially distinct parts of the primary tumor, or from separate FFPE blocks if they were available. Temporal evolution was evaluated with a comparison of primary tumors and recurrences, whenever they were available. Except for the 1p/19q co-deletion, all other CNAs were found to exhibit intratumoral heterogeneity, and varying presence between the primary tumors and recurrences. Contrastingly, the 1p/19q co-deletion was found to be consistently present or absent across all spatially distinct parts of the same tumor, and also across primary tumors and recurrences. This supports the role it is currently considered to have as a defining feature of oligodendrogliomas, and an early event in their development.

Discussion

I will now discuss the presented original publications from today's perspective. This discussion focuses on the three presented bioinformatic solutions: the CanGEM database (chapter 2; Scheinin et al., 2008), the CGHpower sample size tool (chapter 3; Scheinin et al., 2010), and the QDNAseq preprocessing method for copy number detection from shallow WGS data (chapter 4; Scheinin et al., 2014). I will use the names CanGEM, CGHpower, and QDNAseq to refer to these tools and to the original articles that describe them.

Instead of the functionality of these tools, I will focus mainly on the process of developing them. As technology evolves (both in the wet and dry labs), most bioinformatics tools get outdated at some point. Both CanGEM and CGHpower were developed for microarrays, and while they could technically be easily used for DOC-type NGS copy number data, they are probably unlikely to see much of such use. And while currently WES and shallow WGS are excellent cost-efficient solutions to study mutations and copy number, respectively, at some point in the future both might be replaced with deep WGS, if sequencing costs continue their steady decline. For this reason, I will focus on what I have learned from the development process of the presented bioinformatics tools. I hope this provides a more insightful discussion topic than their functionality. Although use of the passive tense is common in scientific writing, here I will frequently use active first-person voice. I feel this better allows me to express views that might be somewhat subjective, but that I hope can be useful considerations for others developing bioinformatics software.

Academic software development

The majority of scientists who develop software as part of their research are primarily self-taught (Hannay et al., 2009; Prabhu et al., 2011). As a result, they often lack knowledge and experience with basic software development practices such as writing maintainable code, using version control and issue trackers, code reviews, unit testing, and task automation (Wilson et al., 2014). I count myself in this group, and believe that inclusion of more training on software development principles in curricula of bioinformatics degree programs could be beneficial and result in better efficiency in development of bioinformatics software.

Wilson et al. (2014) have published a list of recommended best practices for scientific computing. The list is based on their experience in both building scientific software and teaching software development to scientists (Wilson, 2006, 2014), and also on published reports from others. One of the recommendations is to "make incremental changes" based on fast feedback loops and frequent course corrections when needed. This is crucial in the research setting, where the situation and exact requirements can see fast changes (Segal, 2005; Kane et al., 2006; Pitt-Francis et al., 2008; Killcoyne and Boyle, 2009; Pouillon et al., 2011). Two more recommendations from Wilson et al. (2014) will be described in the upcoming discussion.

In addition to the points mentioned above, I would like to discuss the *target audience* of bioinformatics software, because of implementation decisions affected by this choice. The first aspect is whether software is developed for *internal* or *external users*. It is perhaps worth emphasizing that this question is not about usage, but about development. In probably the majority of cases, academic software development is performed to answer a need within the developers' own research group or among close collaborators *(internal users)*. If other research groups have similar needs, they might also be interested in the software and might decide to adopt it for

their use. These, possibly unknown to the developers, people represent *external users*. But in this case, the software was developed for *internal users*, even though *external users* might use it as well. This is always the more straightforward case, because the *internal* needs are known, and are also known to exist. Development for *external users* is more complicated, because the details of the requirements are not necessarily clear. Fulfillment of a perpetual *internal* need is also a good way to ensure the sustainability of academic software development (Broman, 2014).

The second aspect of *target audience* is easiest to discuss by starting from an implementation decision it affects. Software can be roughly divided to graphical user interface (GUI) and command-line interface (CLI) programs. Those with a GUI are commonly deemed easier to use than CLI software, and are therefore seen as more accessible to a wider audience. On the other hand, a GUI program is often less flexible in terms of functionality, and for the developer is typically more laborious to implement, keep up-to-date, and to include support across different operating systems. CLI software is also easier to automate, and is a good fit for reproducible research as processing pipelines can be saved in the form of scripts.

Here, I will make an artificial oversimplification and discuss two possible *target audiences* for bioinformatics software: *bioinformaticians* and *biologists*. Bioinformatics is an interdisciplinary field in the intersection of biosciences, computer science, statistics, mathematics, and engineering. Bioinformaticians are therefore a heterogeneous group with diverse backgrounds, and it is not really possible to define a boundary within the fluid continuum between *bioinformaticians* and *biologists*. But in this text, I will use these terms to refer to bioscientists who are comfortable with CLI programs (and computer programming), and those who are not, respectively. This is naturally an oversimplification, but while not a realistic definition, it conceptually facilitates the discussion that follows. I should also point out that just as end users for software developed for *internal users* might include *external* ones as well, the list of end users can in reality very well contain both *bioinformaticians* and *biologists*, regardless of who the software was primarily developed for. But rather than to attempt to target both of these groups, I find it can be valuable to prioritize one over the other.

One question related to *target audience* and user interface is the division between *method development* for new analytical methods, and mere software *implementations* of existing algorithms. This division is somewhat cloudy; after all, new methods are in practice often described through a software implementation. Nevertheless, this too I believe is a conceptually useful distinction for academic software development.

Bioinformatics software developed for this dissertation

I would now like to discuss the bioinformatic solutions presented in previous chapters from the perspective of software development. Of the presented solutions, QDNAseq is perhaps the most straightforward to discuss. QDNAseq is a DOC *preprocessing* method for the detection of genome-wide chromosomal copy number changes from WGS experiments. It is implemented as a package for R (Ihaka and Gentleman, 1996; R Core Team, 2016), which is a statistical programming language with a CLI user interface. It is distributed through the Bioconductor suite (Huber et al., 2015), which is an open source software project to provide tools for the analysis and comprehension of high-throughput genomic data. It is therefore a clear example of *method development* for *bioinformaticians* (users comfortable with CLI software). QDNAseq was developed to answer a clear need within our own research group. It was therefore developed primarily for *internal* use, but was also made available for *external users* as a part of the Bioconductor suite. Since the user interface is CLI-based, providing additional op-

tions and settings for *external users* adds little overhead.

In addition to the R package, GUI implementations for Chipster (Kallio et al., 2011) and Galaxy (Afgan et al., 2016) were also developed, thus making the method more accessible to a wider audience. But these efforts were separate from the main *method development* process. This separation of mere user interface *implementations* makes it easier to both develop and to maintain the main method itself (Kelly et al., 2009).

Although it is the newest one of the bioinformatic solutions included in this dissertation, QDNAseq is already the most cited one (Table 6.1). Within the Bioconductor software suite (Huber et al., 2015), it is among the "top 20 %" most downloaded packages. And in addition to being used in many prestigious universities and research institutes around the world, it is also in commercial use in the whole-genome copy number assay of the diagnostic company Blueprint Genetics Ltd.

Relative to the other solutions presented in this dissertation, I see QDNAseq as the biggest success. I think part of the reason stems from the implementation choices we made. Most of the existing DOC methods were developed for the analysis of individual samples at a time, and were designed as standalone solutions for the entire *copy number analysis*, including *preprocessing*, *segmentation* and *calling*. In contrast, we designed QDNAseq for the analysis of larger data sets, and also focused it on the *preprocessing* step only, leveraging other Bioconductor packages for *segmentation* and *calling*. One of the recommendations of Wilson et al. (2014) is "don't repeat yourself (or others)", and when standard data structures exist to bridge together multiple methods, this allows them to better focus on their core purpose. When we were evaluating the existing FREEC method (Boeva et al., 2011) while developing QDNAseq, we noticed we could actually improve its performance by replacing the included *segmentation* method (Harchaoui and Lévy-Leduc, 2008) with CBS (Venkatraman and Olshen, 2007). While the output of FREEC does include bin-level data, and thus allows such use of an alternative *segmentation* method, this is not the case for all published methods. For example, the output of readDepth (Miller et al., 2011) only includes *segment*-level data, thus making it impossible to combine its *preprocessing* with any other *segmentation* method.

Development of the CGHpower sample size calculation tool followed a somewhat different path. It was also written in R, but as it was developed to be used by *biologists*, both *internal* and *external*, in the planning stages of new experiments, it was implemented as a web application with a GUI. It is hosted on the web site of the CanGEM database, and can read in data sets directly from the database, thus facilitating analyses for less technical users.

Judging from the number of citations, CGHpower has not generated much interest (Table 6.1). As a power calculation tool, it is intended mainly for the planning stage of experiments, including as a justification for adequate sample size and thus the amount of required funding in grant applications. As such, it might never receive a citation in the resulting scientific article. CGHpower has been used in several grant applications from our research group in Amsterdam, The Tumor Genome Analysis Core, including the grant for the LGG study (Chapter 5; van Thuijl et al., 2014). It is possible other groups might have used the program in their grant applications as well. However, the fact that we found its performance to be sub-optimal for many data sets is perhaps a more probable reason for the lack of interest. Nevertheless, it is difficult to utilize usage measures to judge how successful our implementation decisions were. Also, since the infrastructure had already been built and was maintained for CanGEM, implementing CGHpower as a web application did not require as much resources as it otherwise would have. Still, in retrospect, I think the right approach would

solution	chapter	reference	journal impact factor	citations
CanGEM	2	Scheinin et al., 2008	6.878	20
CGHpower	3	Scheinin et al., 2010	3.029	5
QDNAseq	4	Scheinin et al., 2014	14.630	36

Table 6.1: Number of citations for each of the included bioinformatics solutions, presented as a measure of interest and as a proxy for success. Data obtained from Thomson Reuters Web of Science. Impact factors are for the year of publication, and the number of citations as of September 1, 2017.

have been to release it as an R package, as a CLI tool for *bioinformaticians*. If deemed useful, this could have been supplemented with an additional GUI *implementation*, as was done with QDNAseq.

During the development process, there were a number of situations where a choice could be made between two or more statistical options. Examples include whether to use the Student's t-test (which assumes equal variances) or Welch's t-test (which allows unequal variances), or whether to calculate p-values from the normal or Student's t-distribution. In a CLI program it is easier to offer many such choices to the user as parameters, but we decided to leave these options out from the GUI implementation as we felt at the time that would have caused unnecessary clutter for the user interface.

Regarding CGHpower's sub-optimal performance for many data sets, there are at least two factors that could potentially contribute to the issue. The first one is the assumption of a normal distribution for the median $\log_2$-ratios of *regions*. While in general this text favors the use of *calls* or *call probabilities* over $\log_2$-ratios, here we made this chose because the normal distribution was the only one supported by the power calculation framework (Ferreira and Zwinderman, 2006a). Since then, van Iterson et al. (2013) have extended the framework to also cover the F- and χ^2-distributions. This allows additional experimental designs besides the two-group case, and also offers a more suitable choice for copy number data than the normal distribution we used.

The second factor is the number of data points available to estimate the unknown distribution for *regions* with a real difference between groups. If this number is too small, estimates are unstable and performance suffers. The number of data points available for estimation depends on the total number of *regions*, and also on the proportion of *regions* with a real difference between groups. The examples of both Ferreira and Zwinderman (2006a) and van Iterson et al. (2013) were from mRNA expression, and the dimensionality of the (real and simulated) data sets in the range of thousands or tens of thousands. Contrastingly, the number of *regions* for a copy number experiment is often in the range of hundreds, as was the case for all the (real and simulated) evaluation data sets in Scheinin et al. (2010). Also, since the number of *regions* with a real difference between groups could potentially be small, the power calculation framework might simply have two few data points available to estimate from.

Since the estimation of the power calculation framework works in an asymptotic manner, it actually benefits from high dimensionality. Therefore, while this text in general favors use of *regions* instead of the original features (array elements or sequencing bins), the reduced dimensionality can be a challenge for the power calculation framework.

Of the three presented bioinformatic solutions, CanGEM is perhaps the most interesting discussion topic. It is a centralized database, and the public data sets can be accessed by all users. (Data sets can also be kept private, and read/write access granted on a per user account basis.) This requires it to be implemented on a centralized server. It

was developed to be accessible to a wide audience of *biologists*, both *internal* and *external*, and therefore has a GUI.

Now, nine years after CanGEM was published in the peer-reviewed journal Nucleic Acids Research, it is probably safe to say that it has not been a success. It does work as intended, but what I mean with it not being a success is reflected in the title of the original article: "CanGEM: mining gene copy number changes in cancer". Rather than a description of its functionality, the title is a reflection of CanGEM's ultimate purpose and motivation: to allow data mining studies. However, there is one crucial step from a database that works as intended to one that is a great source for data mining. That step is *data*.

The rationale behind CanGEM was 1) to develop a database that answered the needs of our research group in Helsinki, the Laboratory of Cytomolecular Genetics (CMG): to provide a service where we could store our microarray data, and to be able to query it based on clinical attributes and specific CNAs. Then, 2) if other research groups had similar needs, they might also be interested in using it. And finally, 3) as this could lead to accumulation of data, it would allow novel experiments that utilized data across a number of original experiments.

Today, the CanGEM database contains (as publicly accessible) the raw and processed data of 35 individual studies (Table 6.2). The vast majority of data submissions (28) were performed by people in the CMG group. As collaborators submitted data for six studies, only one submission came from an *external* source. I see this as a clear demonstration that CanGEM did not succeed in its intended purposes 2 and 3. Below, I discuss some alternative implementation decisions we could have made. It is impossible to know if they might have lead to more success, but the aim is to present considerations for future efforts in academic bioinformatics software development.

I would like to start by iterating the usefulness of general software development best practices. During its development process, CanGEM saw one complete rewrite of the entire user interface code, because the original implementation scaled poorly as functionality grew more complex. Also, the initial *copy number analysis* pipeline was based on a combination of R scripts and *segmentation* and *calling* with CGH-Explorer (Lingjaerde et al., 2005). But since CGH-Explorer is a GUI program, only the R parts of the pipeline could be automated, leaving in a step that had to be executed manually. A couple of years after the original publication, CGH-Explorer was replaced with DNAcopy (Venkatraman and Olshen, 2007) and CGHcall (van de Wiel et al., 2007), in order to make the pipeline fully automated. One of the recommendations of Wilson et al. (2014) is to "let the computer do the work" and automate processes that can be automated. This reduces potential for human error, improves reproducibility, and improves sustainability. For these reasons, the *copy number analysis* pipeline of CanGEM should have been designed as fully automated from the beginning. (However, it should perhaps be noted that this particular recommendation of Wilson et al. (2014) should not be over-interpreted in the context of exploratory data analysis, which can require evaluation of intermediate results instead of full automation.)

In addition to the factors that have been mentioned, another reason to develop CanGEM was that data submissions to the existing databases were difficult for *biologists*. They required the original raw data files generated by the image analysis software to be processed to another format, whereas CanGEM was designed to only require the original files. This need could have been answered in another way. Instead of replicating the data warehousing functionality of existing microarray databases, such as GEO (Barrett et al., 2013) or ArrayExpress (Kolesnikov et al., 2015), we could have developed a dedicated *internal* tool to help the submission process to, for example, GEO. And then develop CanGEM as a *copy number anal-*

article	source	submission	arrays
Armengol et al., 2007	internal	internal	10
Atiye et al., 2005	internal	internal	22
Borze et al., 2008	internal	internal	37
Borze et al., 2010	internal	internal	41
Buffart et al., 2008	GEO	internal	2
Cancer Genome Atlas Research Network, 2008	TCGA	collaborator	238
Catrina Ene et al., 2014	internal	internal	20
Ferreira et al., 2008	GEO	internal	25
Heinonen et al., 2008	collaborator	collaborator	32
Järvinen et al., 2008	collaborator	collaborator	61
Junnila et al., 2010a	collaborator	collaborator	31
Junnila et al., 2010b	collaborator	collaborator	20
Kaur et al., 2006a	internal	internal	18
Kaur et al., 2006b	internal	internal	12
Kaur et al., 2007	internal	internal	7
Kaur et al., 2008	internal	internal	19
Koski et al., 2009	collaborator	collaborator	15
Larramendy et al., 2006	internal	internal	4
Lindholm et al., 2007	internal	internal	22
Myllykangas et al., 2008	internal	internal	94
Niini et al., 2010	internal	internal	13
Niini et al., 2011	internal	internal	22
Nymark et al., 2006	internal	internal	20
Savola et al., 2009	internal	internal	47
Siggberg et al., 2011	internal	internal	9
Siggberg et al., 2012	internal	internal	35
Stephan et al., 2008	GEO	internal	25
Szabó et al., 2010	GEO	internal	11
Tap et al., 2011	external	external	46
Tyybäkinoja et al., 2006	internal	internal	4
Usvasalo et al., 2010	internal	internal	103
Vauhkonen et al., 2006a	internal	internal	25
Vauhkonen et al., 2006b	internal	internal	11
Yamashita et al., 2007	GEO	internal	32
Ässämäki et al., 2007	internal	internal	47

Table 6.2: Publicly available data sets in CanGEM.

ysis and querying tool that leveraged raw data stored in GEO. This would have turned CanGEM into more of an indexing tool (instead of raw storage) and thus allowed it to focus more on its core purpose.

Also, instead of requiring sample metadata in a specific file format, CanGEM allowed it to be entered using easy-to-use web forms (while also accepting such a preformatted file for batch submissions). While CanGEM aimed for detailed clinical information and used controlled vocabularies whenever possible, the granularity of sample metadata can be a mixed blessing. Having to answer many detailed questions is much more laborious than a simple free-text field. Also, as with all modern biomedical studies, CNA experiments in cancer research are necessarily a team effort, and can involve clinicians, oncologists, pathologists, geneticists, lab technicians, bioinformaticians, mathematicians, statisticians, and software engineers. While some of the people involved undoubtedly have the information and expertise to answer detailed questions on, for example, the cell-of-origin for a specific tumor type, the person tasked with database submissions might not. If data submissions seem too laborious, it can raise the bar to perform them in the first place. In retrospect, I think we might have aimed too high with CanGEM, and while trying to collect too much information, as a result we received less.

As a service targeted to a wide audience of *external biologists*, CanGEM needed to have a GUI. But with so little data submissions from *external* users (Table 6.2), I think it is relevant to ask if the decision on *target audience* was the right one. If instead of its present form, CanGEM was an indexing tool built on top of raw data stored in GEO and had a CLI for *bioinformaticians* to use, this would have allowed a lot of resources to be spent on, for example, going through existing published data sets and annotating samples with more detailed clinical data. Also, all data input could have been performed *internally*, and only a querying interface for the results built for *external* use, similar to how Kilpinen et al. (2008) did with mRNA expression data from 9,783 samples that represented 68 cancer types, 64 other diseases, and 43 normal human tissue types.

There is also an additional challenge associated with maintaining a centralized service that depends on *external users* for data input. A centralized service requires long-term infrastructure, whereas funding in academia is often based on grants for individual research projects. A common approach is to assign some defined percentage of research grants for longer-term infrastructure needs, possibly on the institution level. While this might secure the necessary resources operationally, a service such as CanGEM that expects *external users* to submit data also has to convince them of long-term sustainability. If a potential user considering submission of their data into the database is worried whether the service will still exist in a few years, they are less likely to go through the trouble. While software as a service (SaaS) has proved to be a good business model for many circumstances and many companies, it is a challenging one for academic software development. Another related challenge is the mobility of people in academia; what happens to the service if the people responsible for its development (often PhD students) or the principal investigator leaves?

With all this auto-criticism of the implementation decisions we made with the CanGEM solution, it is fair to also say that no other database has succeeded either in what we tried to achieve. The Cancer Genome Atlas (TCGA; Cancer Genome Atlas Research Network et al., 2013) is perhaps the closest comparison, and has been a tremendously useful and valuable resource for the cancer research community. It is a centralized repository of processed copy number data across experiments with different cancer types, and is targeted to a wide audience of *biologists* with a GUI on the web. But it was also built for data submissions from within the consortium instead of *external users*, and

also with more solid financial backing than an individual research group.

To summarize the preceding paragraphs, in retrospect I would rather implement CanGEM to 1) leverage existing microarray databases for raw data storage, 2) allow data input only from *internal* users, and possibly through a CLI, and 3) provide a GUI querying interface to the database contents for both *internal* and *external users*.

The discussion above attempts to summarize my personal lessons learned from developing the presented bioinformatic solutions. I hope it can provide at least some useful considerations to people who develop software within academia, including future PhD students in bioinformatics.

Conclusions

Like all modern genomics, the study of genome-wide chromosomal copy number aberrations in cancer relies heavily on computational methods to handle and analyze the large amounts of generated data. Much of the needed software is developed within academic research groups according to their custom needs. This dissertation presents three such bioinformatic solutions. The lessons learned from the development process of these solutions are summarized below.

1. Although software is a crucial component for their work, as software developers most scientists are self-taught. They therefore lack exposure to common software development practices, such as unit tests and code review. These practices have been shown to result in better efficiency in software development, and program code that is less error-prone and easier to maintain. Their value and usefulness is well-known among software engineers, and developers of academic software could benefit from better awareness and utilization of these practices.
2. Software development to answer internal needs is more straightforward than targeting external users. The specific requirements are known more accurately, and it is easier to form fast feedback loops necessary for agile decision-making. Software developed primarily for internal users can be valuable for external users as well, possibly with small modifications. But as a general rule for academic software development, it is perhaps not advisable to be dependent on external users, for example for data submissions.
3. Method development for new analytical algorithms can benefit from modularization. Separation of logically distinct components allows each component to be optimized and reused individually, leading to better overall performance. This is especially true for separation of the program's business logic and user interface.
4. Characteristics of the academic work environment can also have implications for software development. Project-based nature of grant funding and high mobility of key personnel can both have implications for long-term maintenance and support of developed software solutions. These characteristics should be taken into account early on in the development process.

Academic software development could benefit from closer attention to the software development process itself. Providing training for useful software development practices, and more careful considerations of implementation choices and their consequences could result in procedural efficiency, better software, and thus better science.

References

1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, and McVean GA (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**: 56–65.

Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, Grüning B, *et al.* (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res*, **44**: W3–W10.

Alentorn A, van Thuijl HF, Marie Y, Alshehhi H, Carpentier C, Boisselier B, Laigle-Donadey F, Mokhtari K, Scheinin I, Wesseling P, Ylstra B, *et al.* (2014). Clinical value of chromosome arms 19q and 11p losses in low-grade gliomas. *Neuro Oncol*, **16**: 400–408.

Armengol G, Eissa S, Lozano JJ, Shoman S, Sumoy L, Caballín MR, and Knuutila S (2007). Genomic imbalances in Schistosoma-associated and non-Schistosoma-associated bladder carcinoma. An array comparative genomic hybridization analysis. *Cancer Genet Cytogenet*, **177**: 16–19.

Atiye J, Wolf M, Kaur S, Monni O, Bohling T, Kivioja A, Tas E, Serra M, Tarkkanen M, and Knuutila S (2005). Gene amplifications in osteosarcoma-CGH microarray analysis. *Genes Chromosomes Cancer*, **42**: 158–163.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, *et al.* (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res*, **41**: D991–D995.

Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, and Barillot E (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, **27**: 268–269.

Borze I, Juvonen E, Ninomiya S, Jee KJ, Elonen E, and Knuutila S (2010). High-resolution oligonucleotide array comparative genomic hybridization study and methylation status of the RPS14 gene in de novo myelodysplastic syndromes. *Cancer Genet Cytogenet*, **197**: 166–173.

Borze I, Mustjoki S, Mustjokio S, Juvonen E, and Knuutila S (2008). Oligoarray comparative genomic hybridization in polycythemia vera and essential thrombocythemia. *Haematologica*, **93**: 1098–1100.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, *et al.* (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, **29**: 365–371.

Brennan CW, Verhaak RGW, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, *et al.* (2013). The somatic genomic landscape of glioblastoma. *Cell*, **155**: 462–477.

Broman KW (2014). Fourteen years of R/qtl: Just barely sustainable. *J Open Res Softw*, **2**: e11.

Buffart TE, Israeli D, Tijssen M, Vosse SJ, Mrsic A, Meijer GA, and Ylstra B (2008). Across array comparative genomic hybridization: a strategy to reduce reference channel hybridizations. *Genes Chromosomes Cancer*, **47**: 994–1004.

Cancer Genome Atlas Research Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**: 1061–1068.

Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, and Stuart JM (2013). The Cancer Genome Atlas pan-cancer analysis project. *Nat Genet*, **45**: 1113–1120.

Catrina Ene AM, Borze I, Guled M, Costache M, Leen G, Sajin M, Ionica E, Chitu A, and Knuutila S (2014). MicroRNA expression profiles in Kaposi's sarcoma. *Pathol Oncol Res*, **20**: 153–159.

Claus EB and Black PM (2006). Survival rates and patterns of care for patients diagnosed with supratentorial low-grade gliomas: data from the SEER program, 1973–2001. *Cancer*, **106**: 1358–1363.

Douw L, Klein M, Fagel SS, van den Heuvel J, Taphoorn MJ, Aaronson NK, Postma TJ, Vandertop WP, Mooij JJ, Boerman RH, Beute GN, *et al.* (2009). Cognitive and radiological effects of radiotherapy in patients with low-grade glioma: long-term follow-up. *Lancet Neurol*, **8**: 810–818.

Elo LL, Lahti L, Skottman H, Kylaniemi M, Lahesmaa R, and Aittokallio T (2005). Integrating probe-level expression changes across generations of Affymetrix arrays. *Nucleic Acids Res*, **33**: e193.

ENCODE Project Consortium, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, *et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**: 57–74.

Erdem-Eraslan L, Gravendeel LA, de Rooi J, Eilers PHC, Idbaih A, Spliet WGM, den Dunnen WFA, Teepen JL, Wesseling P, Sillevis Smitt PAE, Kros JM, *et al.* (2013). Intrinsic molecular subtypes of glioma are prognostic and predict benefit from adjuvant procarbazine, lomustine, and vincristine chemotherapy in combination with other prognostic factors in anaplastic oligodendroglial brain tumors: a report from EORTC study 26951. *J Clin Oncol*, **31**: 328–336.

Ferreira BI, Alonso J, Carrillo J, Acquadro F, Largo C, Suela J, Teixeira MR, Cerveira N, Molares A, Goméz-López G, Pestaña A, *et al.* (2008). Array CGH and gene-expression profiling reveals distinct genomic instability patterns associated with DNA repair and cell-cycle checkpoint pathways in Ewing's sarcoma. *Oncogene*, **27**: 2084–2090.

Ferreira JA and Zwinderman A (2006a). Approximate sample size calculations with microarray data: an illustration. *Stat Appl Genet Mol Biol*, **5**: Article25.

Ferreira JA and Zwinderman AH (2006b). Approximate power and sample size calculations with the Benjamini-Hochberg method. *Int J Biostat*, **2**: Article8.

Hannay JE, MacLeod C, Singer J, Langtangen HP, Pfahl D, and Wilson G (2009). How do scientists develop and use scientific software? In *Proceedings of the 2009 ICSE Workshop on Software Engineering for Computational Science and Engineering*, SECSE '09, pages 1–8, Washington, DC, USA. IEEE Computer Society.

Harchaoui Z and Lévy-Leduc C (2008). Catching change-points with lasso. *Adv Neural Inform Process Syst*, **20**: 617–624.

Heinonen H, Nieminen A, Saarela M, Kallioniemi A, Klefstrom J, Hautaniemi S, and Monni O (2008). Deciphering downstream gene targets of PI3K/mTOR/p70S6K pathway in breast cancer. *BMC Genomics*, **9**: 348.

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, *et al.* (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods*, **12**: 115–121.

Ihaka R and Gentleman R (1996). R: A language for data analysis and graphics. *J Comp Graph Stat*, **5**: 299–314.

van Iterson M, van de Wiel MA, Boer JM, and de Menezes RX (2013). General power and sample size calculations for high-dimensional genomic data. *Stat Appl Genet Mol Biol*, **12**: 449–467.

Junnila S, Kokkola A, Karjalainen-Lindsberg ML, Puolakkainen P, and Monni O (2010a). Genome-wide gene copy number and expression analysis of primary gastric tumors and gastric cancer cell lines. *BMC Cancer*, **10**: 73.

Junnila S, Kokkola A, Mizuguchi T, Hirata K, Karjalainen-Lindsberg ML, Puolakkainen P, and Monni O (2010b). Gene expression analysis identifies over-expression of CXCL1, SPARC, SPP1, and SULF1 in gastric cancer. *Genes Chromosomes Cancer*, **49**: 28–39.

Järvinen AK, Autio R, Kilpinen S, Saarela M, Leivo I, Grenman R, Mäkitie AA, and Monni O (2008). High-resolution copy number and gene expression microarray analyses of head and neck squamous cell carcinoma cell lines of tongue and larynx. *Genes Chromosomes Cancer*, **47**: 500–509.

Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, Koski M, Käki J, and Korpelainen EI (2011). Chipster: user-friendly analysis software for microarray and other high-throughput data. *BMC Genomics*, **12**: 507.

Kane DW, Hohman MM, Cerami EG, McCormick MW, Kuhlmman KF, and Byrd JA (2006). Agile methods in biomedical software development: a multi-site experience report. *BMC Bioinformatics*, **7**: 273.

Kaur S, Forsman M, Ryhänen J, Knuutila S, and Larramendy ML (2008). No gene copy number changes in Dupuytren's contracture by array comparative genomic hybridization. *Cancer Genet Cytogenet*, **183**: 6–8.

Kaur S, Larramendy ML, Gentile M, Svarvar C, Koivisto-Korander R, Vauhkonen H, Scheinin I, Leminen A, Butzow R, Böhling T, and Knuutila S, *et al.* (2006a). New insights into the cellular pathways affected in primary uterine leiomyosarcoma. *Cancer Genomics Proteomics*, **3**: 347–354.

Kaur S, Larramendy ML, Vauhkonen H, Bohling T, and Knuutila S (2007). Loss of TP53 in sarcomas with 17p12 to approximately p11 gain. A fine-resolution oligonucleotide array comparative genomic hybridization study. *Cytogenet Genome Res*, **116**: 153–157.

Kaur S, Vauhkonen H, Bohling T, Mertens F, Mandahl N, and Knuutila S (2006b). Gene copy number changes in dermatofibrosarcoma protuberans - a fine-resolution study using array comparative genomic hybridization. *Cytogenet Genome Res*, **115**: 283–288.

Kelly D, Hook D, and Sanders R (2009). Five recommended practices for computational scientists who write software. *Comput Sci Eng*, **11**: 48–53.

Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel CV, McCarthy MI, Hide T, *et al.* (2003). eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res*, **13**: 1222–1230.

Killcoyne S and Boyle J (2009). Managing chaos: Lessons learned developing software in the life sciences. *Comput Sci Eng*, **11**: 20–29.

Kilpinen S, Autio R, Ojala K, Iljin K, Bucher E, Sara H, Pisto T, Saarela M, Skotheim RI, Björkman M, Mpindi JP, *et al.* (2008). Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biol*, **9**: R139.

Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, *et al.* (2015). ArrayExpress update–simplifying data submissions. *Nucleic Acids Res*, **43**: D1113–D1116.

Koski T, Lehtonen H, Jee K, Ninomiya S, Joosse S, Vahteristo P, Kiuru M, Karhu A, Sammalkorpi H, Vanharanta S, Lehtonen R, *et al.* (2009). Array comparative genomic hybridization identifies a distinct DNA copy number profile in renal cell cancer associated with hereditary leiomyomatosis and renal cell cancer. *Genes Chromosomes Cancer*, **48**: 544–551.

Larramendy ML, Kaur S, Svarvar C, Bohling T, and Knuutila S (2006). Gene copy number profiling of soft-tissue leiomyosarcomas by array-comparative genomic hybridization. *Cancer Genet Cytogenet*, **169**: 94–101.

Lindholm PM, Salmenkivi K, Vauhkonen H, Nicholson AG, Anttila S, Kinnula VL, and Knuutila S (2007). Gene copy number analysis in malignant pleural mesothelioma using oligonucleotide array CGH. *Cytogenet Genome Res*, **119**: 46–52.

Lingjaerde OC, Baumbusch LO, Liestol K, Glad IK, and Borresen-Dale AL (2005). CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics*, **21**: 821–822.

Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, Ohgaki H, Wiestler OD, Kleihues P, and Ellison DW (2016). The 2016 World Health Organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol*, **131**: 803–820.

Miller CA, Hampton O, Coarfa C, and Milosavljevic A (2011). ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One*, **6**: e16327.

Myllykangas S, Junnila S, Kokkola A, Autio R, Scheinin I, Kiviluoto T, Karjalainen-Lindsberg M, Hollmen J, Knuutila S, Puolakkainen P, and Monni O, *et al.* (2008). Integrated gene copy number and expression microarray analysis of gastric cancer highlights potential target genes. *Int J Cancer*, **123**: 817–825.

Niini T, Lahti L, Michelacci F, Ninomiya S, Hattinger CM, Guled M, Böhling T, Picci P, Serra M, and Knuutila S (2011). Array comparative genomic hybridization reveals frequent alterations of G1/S checkpoint genes in undifferentiated pleomorphic sarcoma of bone. *Genes Chromosomes Cancer*, **50**: 291–306.

Niini T, López-Guerrero JA, Ninomiya S, Guled M, Hattinger CM, Michelacci F, Böhling T, Llombart-Bosch A, Picci P, Serra M, and Knuutila S, *et al.* (2010). Frequent deletion of CDKN2A and recurrent coamplification of KIT, PDGFRA, and KDR in fibrosarcoma of bone–an array comparative genomic hybridization study. *Genes Chromosomes Cancer*, **49**: 132–143.

Nymark P, Wikman H, Ruosaari S, Hollmen J, Vanhala E, Karjalainen A, Anttila S, and Knuutila S (2006). Identification of specific gene copy number changes in asbestos-related lung cancer. *Cancer Res*, **66**: 5737–5743.

Pitt-Francis J, Bernabeu MO, Cooper J, Garny A, Momtahan L, Osborne J, Pathmanathan P, Rodriguez B, Whiteley JP, and Gavaghan DJ (2008). Chaste: using agile programming techniques to develop computational biology software. *Philos Trans A Math Phys Eng Sci*, **366**: 3111–36.

Pouillon Y, Beuken JM, Deutsch T, Torrent M, and Gonze X (2011). Organizing software growth and distributed development: The case of Abinit. *Comput Sci Eng*, **13**: 62–69.

Prabhu P, Jablin TB, Raman A, Zhang Y, Huang J, Kim H, Johnson NP, Liu F, Ghosh S, Beard S, Oh T, *et al.* (2011). A survey of the practice of computational science. In *State of the Practice Reports*, SC '11, pages 19:1–19:12, New York, NY, USA. ACM.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reifenberger J, Reifenberger G, Liu L, James CD, Wechsler W, and Collins VP (1994). Molecular genetic analysis of oligodendroglial tumors shows preferential allelic deletions on 19q and 1p. *Am J Pathol*, **145**: 1175–1190.

Savola S, Klami A, Tripathi A, Niini T, Serra M, Picci P, Kaski S, Zambelli D, Scotlandi K, and Knuutila S (2009). Combined use of expression and CGH arrays pinpoints novel candidate genes in Ewing sarcoma family of tumors. *BMC Cancer*, **9**: 17.

Scheinin I, Ferreira JA, Knuutila S, Meijer GA, van de Wiel MA, and Ylstra B (2010). CGHpower: exploring sample size calculations for chromosomal copy number experiments. *BMC Bioinformatics*, **11**: 331–340.

Scheinin I, Myllykangas S, Borze I, Böhling T, Knuutila S, and Saharinen J (2008). CanGEM: mining gene copy number changes in cancer. *Nucleic Acids Res*, **36**: D830–D835.

Scheinin I, Sie D, Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF, van Essen HF, Eijk PP, Rustenburg F, Meijer GA, Reijneveld JC, *et al.* (2014). DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res*, **24**: 2022–2032.

Segal J (2005). When software engineers met research scientists: a case study. *Empir Softw Eng*, **10**: 517–536.

Siggberg L, Ala-Mello S, Linnankivi T, Avela K, Scheinin I, Kristiansson K, Lahermo P, Hietala M, Metsähonkala L, Kuusinen E, Laaksonen M, *et al.* (2012). High-resolution SNP array analysis of patients with developmental disorder and normal array CGH results. *BMC Med Genet*, **13**: 84.

Siggberg L, Peippo M, Sipponen M, Miikkulainen T, Shimojima K, Yamamoto T, Ignatius J, and Knuutila S (2011). 9q22 deletion–first familial case. *Orphanet J Rare Dis*, **6**: 45.

Smyth GK and Speed T (2003). Normalization of cDNA microarray data. *Methods*, **31**: 265–273.

Stephan EA, Chung TH, Grant CS, Kim S, Von Hoff DD, Trent JM, and Demeure MJ (2008). Adrenocortical carcinoma survival rates correlated to genomic copy number variants. *Mol Cancer Ther*, **7**: 425–431.

Szabó PM, Tamási V, Molnár V, Andrásfalvy M, Tömböl Z, Farkas R, Kövesdi K, Patócs A, Tóth M, Szalai C, Falus A, *et al.* (2010). Meta-analysis of adrenocortical tumour genomics data: novel pathogenic pathways revealed. *Oncogene*, **29**: 3163–3172.

Tap WD, Eilber FC, Ginther C, Dry SM, Reese N, Barzan-Smith K, Chen HW, Wu H, Eilber FR, Slamon DJ, and Anderson L, *et al.* (2011). Evaluation of well-differentiated/dedifferentiated liposarcomas by high-resolution oligonucleotide array-based comparative genomic hybridization. *Genes Chromosomes Cancer*, **50**: 95–112.

van Thuijl HF, Scheinin I, Sie D, Alentorn A, van Essen HF, Cordes M, Fleischeuer R, Gijtenbeek AM, Beute G, van den Brink WA, Meijer GA, *et al.* (2014). Spatial and temporal evolution of distal 10q deletion, a prognostically unfavorable event in diffuse low-grade gliomas. *Genome Biol*, **15**: 471–483.

van Thuijl HF, Ylstra B, Würdinger T, van Nieuwenhuizen D, Heimans JJ, Wesseling P, and Reijneveld JC (2012). Genetics and pharmacogenomics of diffuse gliomas. *Pharmacol Ther*, **137**: 78–88.

Tyybäkinoja A, Saarinen-Pihkala U, Elonen E, and Knuutila S (2006). Amplified, lost, and fused genes in 11q23-25 amplicon in acute myeloid leukemia, an array-CGH study. *Genes Chromosomes Cancer*, **45**: 257–264.

Usvasalo A, Elonen E, Saarinen-Pihkala UM, Räty R, Harila-Saari A, Koistinen P, Savolainen ER, Knuutila S, and Hollmén J (2010). Prognostic classification of patients with acute lymphoblastic leukemia by using gene copy number profiles identified from array-based comparative genomic hybridization data. *Leuk Res*, **34**: 1476–1482.

Vauhkonen H, Vauhkonen M, Sajantila A, Sipponen P, and Knuutila S (2006a). Characterizing genetically stable and unstable gastric cancers by microsatellites and array comparative genomic hybridization. *Cancer Genet Cytogenet*, **170**: 133–139.

Vauhkonen H, Vauhkonen M, Sajantila A, Sipponen P, and Knuutila S (2006b). DNA copy number aberrations in intestinal-type gastric cancer revealed by array-based comparative genomic hybridization. *Cancer Genet Cytogenet*, **167**: 150–154.

Venkatraman ES and Olshen AB (2007). A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**: 657–663.

Wesseling P, van den Bent M, and Perry A (2015). Oligodendroglioma: pathology, molecular mechanisms and markers. *Acta Neuropathol*, **129**: 809–827.

van de Wiel MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, and Ylstra B (2007). CGHcall: calling aberrations for array CGH tumor profiles. *Bioinformatics*, **23**: 892–894.

van de Wiel MA and van Wieringen WN (2007). CGHregions: dimension reduction for array CGH data with minimal information loss. *Cancer Informatics*, **3**: 55–63.

Wilson G (2006). Software Carpentry: Getting scientists to write better code by making them more productive. *Comput Sci Eng*, **8**: 66–69.

Wilson G (2014). Software Carpentry: lessons learned. *F1000Res*, **3**: 62.

Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, Haddock SHD, Huff KD, Mitchell IM, Plumbley MD, Waugh B, *et al.* (2014). Best practices for scientific computing. *PLoS Biol*, **12**: e1001745.

Yamashita Y, Minoura K, Taya T, Fujiwara Si, Kurashina K, Watanabe H, Choi YL, Soda M, Hatanaka H, Enomoto M, Takada S, *et al.* (2007). Analysis of chromosome copy number in leukemic cells by different microarray platforms. *Leukemia*, **21**: 1333–1337.

Yan H, Parsons DW, Jin G, McLendon R, Rasheed BA, Yuan W, Kos I, Batinic-Haberle I, Jones S, Riggins GJ, Friedman H, *et al.* (2009). IDH1 and IDH2 mutations in gliomas. *N Engl J Med*, **360**: 765–773.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L, Girón CG, *et al.* (2016). Ensembl 2016. *Nucleic Acids Res*, **44**: D710–D716.

Zhang Z, Schwartz S, Wagner L, and Miller W (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol*, **7**: 203–214.

Ässämäki R, Sarlomo-Rikala M, Lopez-Guerrero JA, Lasota J, Andersson LC, Llombart-Bosch A, Miettinen M, and Knuutila S (2007). Array comparative genomic hybridization analysis of chromosomal imbalances and their target genes in gastrointestinal stromal tumors. *Genes Chromosomes Cancer*, **46**: 564–576.

Full list of publications

In reverse chronological order.

1. Mäki-Nevala S, Sarhadi VK, Knuuttila A, Scheinin I, Ellonen P, Lagström S, Rönty M, Kettunen E, Husgafvel-Pursiainen K, Wolff H, Knuutila S (2016) **Driver gene and novel mutations in asbestos-exposed lung adenocarcinoma and malignant mesothelioma detected by exome sequencing.** *Lung* **194:** 125–135
2. Prabowo AS[1], van Thuijl HF[1], Scheinin I, Sie D, van Essen HF, Iyer AM, Spliet WG, Ferrier CH, van Rijen PC, Veersema TJ, Thom M, Schouten-van Meeteren AY, Reijneveld JC, Ylstra B, Wesseling P[2], Aronica E[2] (2015) **Landscape of chromosomal copy number aberrations in gangliogliomas and dysembryoplastic neuroepithelial tumours.** *Neuropathology and Applied Neurobiology* **41:** 743–755
3. van Thuijl HF[1], Scheinin I[1], Sie D, Alentorn A, van Essen HF, Cordes M, Fleischeuer R, Gijtenbeek AM, Beute G, van der Brink WA, Meijer GA, Havenith M, Idbaih A, Hoang-Xuan K, Mokhtari K, Verhaak RGW, van der Valk P, van de Wiel MA, Heimans JJ, Aronica E, Reijneveld JC, Wesseling P, Ylstra B (2014) **Spatial and temporal evolution of distal 10q deletion, a prognostically unfavorable event in diffuse low-grade gliomas.** *Genome Biology* **15:** 471–483.
4. Scheinin I[1], Sie D[1], Bengtsson H, van de Wiel MA, Olshen AB, van Thuijl HF, van Essen HF, Eijk PP, Rustenburg F, Meijer GA, Reijneveld JC, Wesseling P, Pinkel D, Albertson DG, Ylstra B (2014) **DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly.** *Genome Research* **24:** 2022–2032.
5. Sarhadi VK, Lahti L, Scheinin I, Ellonen P, Kettunen E, Serra M, Scotlandi K, Picci P, Knuutila S (2014) **Copy number alterations and neoplasia-specific mutations in MELK, PDCD1LG2, TLN1, and PAX5 at 9p in different neoplasias.** *Genes Chromosomes Cancer* **53:** 579–588.
6. Alentorn A, van Thuijl HF, Marie Y, Alshehhi H, Carpentier C, Boisselier B, Laigle-Donadey F, Mokhtari K, Scheinin I, Wesseling P, Ylstra B, Capelle L, Hoang-Xuan K, Sanson M, Delattre JY, Reijneveld JC, Idbaih A (2014) **Clinical value of chromosome arms 19q and 11p losses in low-grade gliomas.** *Neuro-Oncology* **16:** 400–408.
7. Sarhadi VK[1], Lahti L[1], Scheinin I, Tyybäkinoja A, Savola S, Usvasalo A, Räty R, Elonen E, Ellonen P, Saarinen-Pihkala UM, Knuutila S. (2013) **Targeted resequencing of 9p in acute lymphoblastic leukemia yields concordant results with array CGH and reveals novel genomic alterations.** *Genomics* **102:** 182–188.
8. Niini T, Scheinin I, Lahti L, Savola S, Mertens F, Hollmén J, Böhling T, Kivioja A, Nord KH, Knuutila S (2012) **Homozygous deletions of cadherin genes in chondrosarcoma - an array CGH study.** *Cancer Genetics* **205:** 588–593.

9. Siggberg L, Ala-Mello S, Linnankivi T, Avela K, Scheinin I, Kristiansson K, Lahermo P, Hietala M, Metsähonkala L, Kuusinen E, Laaksonen M, Saarela J, Knuutila S (2012) **High-resolution SNP array analysis of patients with developmental disorder and normal array CGH results.** *BMC Medical Genetics* **13:** 84–93.

10. Kallio AM[1], Tuimala JT[1], Hupponen T, Klemelä P, Gentile M, Scheinin I, Koski M, Käki J, Korpelainen EI (2011) **Chipster: user-friendly analysis software for microarray and other high-throughput data.** *BMC Genomics 12:* 507–520.

11. Borze I, Scheinin I, Siitonen S, Elonen E, Juvonen E, Knuutila S (2011) **miRNA expression profiles in myelodysplastic syndromes reveal Epstein-Barr virus miR-BART13 dysregulation.** *Leukemia & Lymphoma* **52:** 1567–1573.

12. Scheinin I, Ferreira JA, Knuutila S, Meijer GA, van de Wiel MA, Ylstra B (2010) **CGHpower: exploring sample size calculations for chromosomal copy number experiments.** *BMC Bioinformatics* **11:** 331–340.

13. Myllykangas S1, Junnila S1, Kokkola A, Autio R, Scheinin I, Kiviluoto T, Karjalainen-Lindsberg M-L, Hollmén J, Knuutila S, Puolakkainen P, Monni O (2008) **Integrated gene copy number and expression microarray analysis of gastric cancer highlights potential target genes.** *International Journal of Cancer* **123:** 817–825.

14. Scheinin I, Myllykangas S, Borze I, Bohling T, Knuutila S, Saharinen J (2008) **CanGEM: mining gene copy number changes in cancer.** *Nucleic Acids Research* **36:** D830–D835.

15. Kaur S, Larramendy ML, Gentile M, Svarvar C, Koivisto-Korander R, Vauhkonen H, Scheinin I, Leminen A, Bützow R, Böhling T, Knuutila S (2006) **New insights into the cellular pathways affected in primary uterine leiomyosarcoma.** *Cancer Genomics & Proteomics* **3:** 347–354.

[1] Shared first authors.
[2] Shared last authors.

Acknowledgments

This work was carried out at the Department of Pathology, University of Helsinki during 2005–2011, and at the Department of Pathology, VU University Medical Center during 2011–2014. I would like to express my gratitude to all of those who have contributed towards its completion. Now former heads of department, **Veli-Pekka Lehto** and **Gerrit Meijer**, are thanked for providing the research facilities.

First and foremost, I would like to thank my supervisors **Sakari Knuutila** and **Bauke Ylstra** for all their guidance and support. I am grateful for everything from securing funding for my work, to the valuable lessons I have learned on cancer and science in general, writing and presentation skills, project management, and about the realities of the academic research world. I am also grateful for their great company and passionate discussions about tour skating.

During the CanGEM project, I had the pleasure to be also supervised by **Juha Saharinen**. He left a lasting impression on a young bioinformatics PhD student, as I could never make up my mind on whether he was more of an expert of biochemistry and molecular biology, or of computers and programming, as he seemed to always excel in both areas. **Samuel Myllykangas** is thanked for his role during the early parts of this work that, for lack of a better word, I would define as an academic older brother. He took me under his wing and guided me through the peculiarities of academia.

Mark van de Wiel is thanked especially for his role as a co-promoter and his input on the text of this dissertation, and also for all of our highly valuable discussions on mathematics and statistics. **Massimiliano Gentile** is thanked for great discussions not only regarding statistics and data analysis, but also cycling and the great outdoors in general. He is not a co-author on any of the articles that are included in this dissertation, but over the course of my PhD studies we had so many and so educative professional discussions that he deserves a special mention here. **Henrik Bengtsson** is thanked for mentorship on both statistics and software development. From him I have learned so much that it feels strange that we have actually never met in person. I have probably never worked with anyone who is as clear, unambiguous, and thoughtful communicator over email. **José Ferreira**, **Adam Olshen**, and **Wessel van Wieringen** are also thanked for sharing their mathematical and statistical knowledge with me.

Other domain experts I have had the privilege to work with include **Tom Böhling** (pathology), **Pieter Wesseling** (pathology), **Jaap Reijneveld** (neuro-oncology), and **Jukka Vakkila** (pediatrics and immunology). I am grateful for the expertise they have provided towards this work. **Donna Albertson** and **Dan Pinkel** are thanked for both their expertise and also for our numerous Skype discussions that taught me a lot about the importance of relentlessly going after the little details that just are not right yet. **Ioana Borze** is thanked for her role as the main data submitter of the CanGEM database, and **Juri Ahokas**, **Tomi Simonen**, **Teemu Perheentupa**, **Jani Heikkinen**, and **Olle Hansson** for their work in providing IT infrastructure maintenance and support.

For direct contributions related to this dissertation, I would like to thank the official reviewers, **Pieter Wesseling**, **Lodewyk Wessels**, **Sanne Abeln**, and especially **Matti Nykter** and **Laura Elo**, who performed the task for both universities and gave thoughtful comments on the text. My uncle **Henry Scheinin** is thanked for designing the book cover, and **Jos Poell** and **Erik van Dijk** for their help with the language of the abstracts. I would also like to express my gratitude to all co-authors of the included

publications, and to everyone whose contributions have been crucial in order to generate the data for me to analyze.

My fellow (former) PhD students in bioinformatics, **Daoud Sie**, **Oscar Krijgsman**, **Leo Lahti**, are thanked for all their peer support and company. In a work environment where one is surrounded mostly by professionals of parallel fields (molecular biologists, medical doctors, lab technicians, etc), it is great to have the company and support of one's peers. They are people I have learned so much from, and who I could trust to always understand me.

For the cotutelle agreement between the two universities, I would like to thank the now former rector magnificus **Frank van der Duyn Schouten**, now former dean **Wim Stalman**, and dean **Risto Renkonen**. For their help in preparation of the agreement, I would like to thank **Esko Koponen**, **Pentti Tienari**, **Jan Bekker**, **Peter Brasik**, **Pauline Mulligen**, and especially **Katja Juntunen** whose contribution to the entire process was absolutely crucial. **Tarja Nieminen**, **Janneke van Denderen**, **Milja Tikkanen**, **Pirjo Pennanen**, **Mari Siltala**, and **Corien Soutendijk** are thanked for their help with other official matters.

I would like to thank all the people with whom I have had the pleasure to share an office, **Daniëlle Israeli**, **Hinke van Thuijl**, **Matias Mendeville**, **Dirk van Essen**, **Paul Eijk**, **Serge Smeets**, **Neda Mosakhani**, **Mohamed Guled**, **Tommi Ripatti**, and **Juha Knuuttila**. With them I have had numerous interesting and entertaining conversations, and I could always count on them to brighten up a perhaps otherwise gloomy morning.

I would also like to thank all my former colleagues who have not yet been mentioned: **Martijn Cordes**, **Josien Haan**, **François Rustenburg**, **Linda Forsström**, **Sanna Heino**, **Kowan Jee**, **Sippy Kaur**, **Eeva Kettunen**, **Pamela Lindholm**, **Tuija Lundan**, **Satu Mäki-Nevala**, **Tarja Niini**, **Sinshuke Ninomiya**, **Penny Nymark**, **Fabricio Passador-Santos**, **Salla Ruosaari**, **Virinder Kaur Sarhadi**, **Suvi Savola**, **Katja Tuononen**, **Anne Tyybäkinoja**, **Anu Usvasalo**, **Hanna Vauhkonen**, and **Tiina Wirtanen**. It has been a pleasure to part of the great work communities these people have formed together.

For their support throughout the process of working on this dissertation, I would like to thank all my **friends**. I would also like to give a special salute to the crew of MS *Isla*, as they made it possible for me to focus on writing by isolating myself to our island cottage on the Finnish Archipelago Sea. And finally, I want to express my warmest gratitude to my **family**, my mother **Helvi**, my father **Martin** and his wife **Margot**, my sister **Mirjam**, and especially my girlfriend **Heidi**, for all their love and continuous support. They have always been there for me, made sure I had everything I needed, endured me when the work might have taken its toll on me, and allowed me to ignore other matters and focus on writing during critical moments. I could not have done this without you.

Thank you.

Helsinki, September 2017,